

A Unified Paradigm of Organized Complexity and Semantic Information Theory

Tatsuaki Okamoto

NTT

3-9-11 Midori-cho, Musashino-shi, Tokyo, 180-8585 Japan

okamoto.tatsuaki@lab.ntt.co.jp

(Dated: August 3, 2016)

One of the most fundamental problems in science is to define *quantitatively* the complexity of organized matters, i.e., *organized complexity*. Although many measures have been proposed toward this aim in previous decades, there is no agreed upon definition. This paper presents a new quantitative definition of organized complexity. In contrast to existing measures such as the Kolmogorov complexity, logical depth, effective complexity, and statistical complexity, this new definition *simultaneously* captures the three major features of complexity: computational (similar to logical depth), descriptive (similar to the Kolmogorov complexity and effective complexity) and distributional (similar to statistical complexity). In addition, the proposed definition is computable and can measure both probabilistic and deterministic forms of objects in a unified manner. The proposed definition is based on circuits rather than Turing machines and ϵ -machines. We give several criteria required for organized complexity measures and show that the proposed definition satisfies all of them for the first time.

We then apply this quantitative definition to formulate a *semantic information theory*. We present the first formal definition of a *semantic information amount*, which is the core concept of the semantic information theory, that is based only on concretely defined notions. Previous semantic information theories defined this amount under some a priori information which is not concretely specified. We then unveil several fundamental properties in the semantic information theory, e.g., a semantic source coding theorem, semantic channel coding theorem, and effectiveness coding theorem. Although the semantic information theory has a long history of research going back more than six decades, there has been no study on its relation to organized complexity. This paper offers the first unified paradigm of organized complexity and semantic information theory.

I. INTRODUCTION

A. Background

Around seven decades ago, an American scientist, Warren Weaver, classified scientific problems into three classes: problems of *simplicity*, problems of *disorganized complexity*, and problems of *organized complexity* [40]. For example, the classical dynamics can be used to analyze and predict the motion of a few ivory balls as they move about on a billiard table. This is a typical problem of simplicity. Imagine then, a large billiard table with millions of balls rolling over its surface, colliding with one another and with the side rails. Although to be sure the detailed history of one specific ball cannot be traced, statistical mechanics can analyze and predict the average motions. This is a typical problem of disorganized complexity. Problems of organized complexity, however, deal with features of organization such as living things, ecosystems, and artificial things. Here, cells in a living thing are interrelated into an organic whole in their positions and motions, whereas the balls in the above illustration of disorganized complexity are distributed in a helter-skelter manner.

In the tradition of Lord Kelvin, the quantitative definition of complexity is the most fundamental and important notion in problems of complexity.

“I often say that when you can measure what you are speaking about, and express it in numbers, you know

something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of science, whatever the matter may be.”

Lord Kelvin, 1883

The quantitative definition of *disorganized complexity* of physical systems has been established to be *entropy*, which is defined in thermodynamics and statistical mechanics. In a similar manner, disorganized complexity of information sources (distributions) can be quantitatively defined by Shannon entropy [35].

In contrast, there is no agreed upon quantitative definition of *organized complexity*. The difficulty comes from the notion that organized complexity could be greatly dependent on our senses or that the objects of organized complexity like living things, ecosystems, and artificial things may be recognized only by intelligent organisms like human beings, that is to say, it is vastly different from the measures of disorganized complexity such as entropy and Shannon entropy which simply quantify the randomness of the objects.

We may therefore wonder whether such sensory and vague things can be rigorously defined in a unified manner covering various living things to artificial things. Many investigations nonetheless have been pursued toward this aim in the last decades, e.g., *logical depth*

by Bennett [4], *effective complexity* by Gel-Man [15–17], *thermodynamic depth* by Lloyd and Pagels [28], *effective measure complexity* by Grassberger [18], and *statistical complexity* by Crutchfield et al. [8–10, 36], although no existing measure has been agreed on in the field. [24, 25]. In Section II C, we explain our understanding on why no existing measure is satisfactory to be agreed upon.

The quantitative definition of complexity of an object is essentially related to the amount of information that the object possesses. For example, Shannon entropy, which is the quantitative definition of disorganized complexity of an information source, was introduced to define the amount of information of a source in the sense of Shannon’s information theory [35].

One year after Shannon introduced his information theory (and Weaver published the aforementioned article [40]), Weaver proposed that there are three levels of information and communication problems [41]:

- Level A: How accurately can the symbols of communication be transmitted? (The technical problem.)
- Level B: How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
- Level C: How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Interestingly, the classes of scientific problems classified by Weaver [40] are closely related to the above-mentioned three levels of information problems. The *disorganized* complexity measures, e.g., Shannon entropy, are related to the Level A problem, *technical or syntactic* problem, e.g., Shannon information theory, and the *organized* complexity measures should be related to the Level B and C problems, i.e., *semantic and effectiveness* problems.

There have been many studies on the Level B (semantic) problem spanning more than six decades in terms of the semantic information theory [2, 3, 5, 11, 13, 14, 19–21, 23, 29, 31, 38, 42], but no existing work has been recognized as a standard theory. In Section III A, we show a fundamental problem common among the existing works for Level B that forms the basis of our understanding of why no existing work could be a standard theory. Roughly, all existing work assumes some a priori information which is not concretely specified, where informal observation might be possible but no formal result could be achieved rigorously without concrete specification of such a priori information. In addition, no observation has been presented in literature on the relation between the semantic (Level B) problem and the organized complexity. To the best of our knowledge, no study has been conducted seriously on the Level C (effectiveness) problem.

B. Contribution

This paper presents a new quantitative definition of organized complexity. In contrast to existing measures such as Kolmogorov complexity, logical depth, effective complexity and statistical complexity, this new definition simultaneously captures the three major features of complexity: computational feature (similar to logical depth), descriptive feature (similar to Kolmogorov complexity and effective complexity) and distributional feature (similar to statistical complexity). In addition, the proposed definition is computable and can measure both probabilistic and deterministic forms of objects in a unified manner. The proposed definition is based on *circuits* [34, 39] rather than Turing machines [4, 15–17, 27, 34] and ϵ -machines [8–10, 36]. Our new measure is given by the shortest size of a stochastic finite-state automaton form of circuit, *oc-circuit*, for simulating the object. Here note that, given an object, the shortest size of an oc-circuit to simulate the object is computable and that the size of an oc-circuit can capture the computational, descriptive and distributional features of complexity of the object. We give several criteria required for organized complexity measures and show that the proposed definition is the first that satisfies all of the requirements.

We then present the first semantic information theory for the Level B (semantic) problem that overcomes the fundamental problem common among all previous works. That is, the proposed semantic information theory is constructed only on concretely defined notions. This theory is based on the proposed organized complexity measure. We then unveil several fundamental properties in the semantic information theory, e.g., a semantic source coding theorem and semantic channel coding theorem. Moreover, this paper, for the first time, develops a theory for the effectiveness (Level C) problem, which is also constructed on our organized complexity measure. In other words, we clarify the relationship of organized complexity with the semantic and effectiveness (Level B and C) problems of information and communication.

Thus, this paper presents the first unified paradigm for the organized complexity and the semantic information theory that covers the semantic and effectiveness problems.

C. Notations

The sets of natural, rational, and real numbers are denoted by \mathbb{N} , \mathbb{Q} , and \mathbb{R} , respectively. The set of n -bit strings is denoted by $\{0, 1\}^n$ ($n \in \mathbb{N}$), $\{0, 1\}^* := \bigcup_{n \in \mathbb{N}} \{0, 1\}^n$, and the null string (0-bit string) is denoted by λ . When $x \in \{0, 1\}^*$, $|x|$ denotes the bit length of x . When $a, b \in \mathbb{R}$, $[a, b]$ denotes set $\{x \mid x \in \mathbb{R}, a \leq x \leq b\} \subset \mathbb{R}$. When $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the smallest integer greater than or equal to x .

When x is a variable and y is a value or $x := y$ denotes that x is substituted or defined by y . A probability

distribution over $\{0,1\}^n$ is $\{(a, p_a) \mid a \in \{0,1\}^n, p_a \in [0,1], \sum_{a \in \{0,1\}^n} p_a = 1\}$. When A is a probability distribution, or the source (machinery) of the distribution, $a \stackrel{R}{\leftarrow} A$ denotes that element $a \in \{0,1\}^n$ is randomly selected from A according to its probability distribution. When A is a set, $a \stackrel{U}{\leftarrow} A$ denotes that a is randomly selected from A with a uniform distribution.

When X and Y are two distributions, the statistical distance of X and Y , $\text{SD}(X, Y)$, is defined by $\frac{1}{2} \cdot \sum_{\alpha \in \{0,1\}^*} |\Pr[\alpha \stackrel{R}{\leftarrow} X] - \Pr[\alpha \stackrel{R}{\leftarrow} Y]|$, and $X \stackrel{\delta}{\approx} Y$ denotes that $\text{SD}(X, Y)$ is bounded by δ . Then we say X and Y are statistically δ -close.

When Y is a distribution, $(Y)_n$ denotes the n -bit restriction of Y , i.e., $(Y)_n$ is a distribution over $\{0,1\}^n$ and $\Pr[y \stackrel{R}{\leftarrow} (Y)_n] = \sum_{z \in \{0,1\}^*} \Pr[(y, z) \stackrel{R}{\leftarrow} Y]$.

When S is a set, $\#S$ denotes the number of elements of S .

II. ORGANIZED COMPLEXITY MEASURE

A. Objects

The existing complexity measures can be categorized in two classes. One is a class of measures whose objects are *deterministic strings*, and the objects of the other class of measures are *probability distributions*. As for the traditional complexity measures, the Kolmogorov complexity is categorized in the former and (Shannon) entropy is in the latter. Among the above-mentioned organized complexity measures, logical depth and effective complexity are in the former, and thermodynamic depth, effective measure complexity, and statistical depth are in the latter.

Which is more appropriate as the objects of organized complexity?

The objects of complexity are everything around us, stars and galaxies in space, living things, ecosystems, artificial things, and human societies. The existence of everything can be recognized by us only through observations. For example, the existence of many things are observed through devices such as the telescope, microscope, various observation apparatus and electronic devices. We can take things around us directly into our hands and sense them, but they are also recognized by our brains as electronic nerve signals transmitted from the sensors of our five senses through the nervous system. That is, all objects of complexity are recognized as the result of observations by various apparatus and devices including the human sensors of our five senses.

Since the micro world is governed by quantum mechanics, observed values are determined in a probabilistic manner. This is because observed values (data) obtained when observing micro phenomena (quantum states) in quantum mechanics are randomly selected according to a

certain probability distribution corresponding to a quantum state (e.g., entangled superposition).

How then, are observed values in macro phenomena? For example, if some sort of radio signals are received and measured, they would almost certainly be accompanied by noise. There are various reasons why noise becomes mixed in with signals, and one of them is thought to be the probabilistic phenomena of electrons, thermal noise. Similarly, various types of noise will be present in the data obtained when observing distant astronomical bodies. This can be caused by the path taken by the light (such as through the atmosphere) and by factors associated with the observation equipment.

Even in the case of deterministic physical phenomena, chaos theory states that fluctuations in initial conditions can lead to diverse types of phenomena that behave similar to those of random systems. In short, even a deterministic system can appear to be a quasi-probabilistic system. But even a system of this type can become a true (non-quasi) probabilistic system if initial conditions fluctuate due to some noise, e.g., thermal noise. There are also many cases in which a quasi-probabilistic system associated with chaos cannot be distinguished from a true probabilistic system depending on the precision of the observation equipment. Here, even quasi-probabilistic systems may be treated as true probabilistic systems.

Thus, when attempting to give a quantitative definition of the complexity of observed data, the source of those observed data would be a probability distribution and the observed data themselves would be values randomly selected according to that distribution. If we now consider the complexity of a phenomenon observed using certain observation equipment, the object of this complexity should not be the observed data selected by chance from the source but rather the probability distribution itself corresponding to the source of the observed data.

It is known that some parts of genome patterns appear randomly distributed over a collection of many samples (over generations). Here, we can suppose a source (probability distribution) of genome patterns, from which each genome pattern is randomly selected. Also in this case, the object of complexity should not be each individual genome pattern but rather the source (probability distribution) of the genome patterns.

Therefore, hereafter in this paper, we consider that an object of complexity is a *probability distribution*. Here note that a deterministic string can be considered to be a very special case of a probability distribution (where only a value occurs with probability 1 and the others with 0).

How can we determine a source or probability distribution from observed data? It has been studied as the *model selection theory* in statistics and information theory, e.g., AIC (Akaike's information criterion) by Akaike [1] and MDL (minimum description length) by Rissanen [32, 33], given a collection of data, to find the most likely and succinct model (source, i.e., probability distribution) of the

data. In this paper, however, it is outside the scope, i.e., we do not consider how to find such a source from a collection of observed data (through the model selection theory). We here suppose that a source (probability distribution) is given as an object of organized complexity, and focus on how to define quantitatively the complexity of such a given source.

There are roughly two types of observed data, one type is data observed at a point in time and the other is time series data. Genome pattern data are an example of the former, and data obtained from an observation apparatus for a certain time period are an example of the latter. In any case, without loss of generality, we here assume that observed data x are bounded and expressed in binary form, i.e., $x \in \{0, 1\}^n$ for some $n \in \mathbb{N}$, since any physically observed data have only finite precision (no infinite precision). Then the *source* of the observed data, X , which is an object of organized complexity in this paper, is a probability distribution over $\{0, 1\}^n$ for some $n \in \mathbb{N}$ such that $X := \{(x, p_x) \mid x \in \{0, 1\}^n, 0 \leq p_x \leq 1, \sum_{x \in \{0, 1\}^n} p_x = 1\}$.

B. Criteria

We describe our attempt to define quantitatively the organized complexity. To begin with, let us consider the following example. We give a chimpanzee a computer keyboard and prompt the chimpanzee to hit the keys freely resulting in the output of a string of characters. Let us assume an output of 1000 alphabetical characters. At the same time, we select a string of 1000 characters from one of Shakespeare's plays. Naturally, the character string input by the chimpanzee is gibberish possessing no meaning, which undoubtedly makes it easy for us to distinguish that string from a portion of a Shakespearean play.

Is there a way, however, to construct a mathematical formulation of the difference between these two strings that we ourselves can easily tell apart? Why is it so easy for us to make a distinction between these two strings? The answer is likely that the chimpanzee's string is simply random (or disorganized) and meaningless to us while Shakespeare's string is highly organized and meaningful. In short, if we can mathematically define the amount of organized complexity (or meaningful information), we should be able to make a distinction between these two strings.

What then are the sources of the observed data, the chimpanzee's string and Shakespeare's string. Let us return to the source of the chimpanzee's string creation without thinking of it as simply a deterministic string. Here, for the sake of simplicity, we suppose that the scattered hitting of keys by the chimpanzee is the same as a random selection of hit keys. At this time, the source of the chimpanzee's string is the probability distribution in which any particular 1000-character string can be randomly selected from all possible 1000-character strings

with equal probability.

What, then, would be the source of Shakespeare's string? We can surmise that, when Shakespeare wrote down this particular 1000-character string, a variety of expressions within his head would have been candidates for use, and that the 1000-character string used in the play would have finally been selected from those candidates with a certain probability. The candidates selected must certainly be connected by complex semantic relationships possessed by English words. Accordingly, the source of Shakespeare's string must be the complex probability distribution of candidate expressions connected by complex semantic relationships. For example (Case 1), candidate expression 1 has the probability of 0.017, candidate expression 2 has the probability of 0.105, ..., candidate expression 327 has the probability of 0.053 and the other expressions have the probability of 0, i.e., hundreds of candidate expressions occurred in his head consciously or unconsciously and finally one of them was randomly chosen according to the distribution. As more simplified cases, Case 2 is where candidate expression 1 has the probability of $2/7$, candidate expression 2 has the probability of $5/7$ and the others have the probability of 0, i.e., only two candidate expressions occurred in his head and finally one of them was randomly chosen. Case 3 is where only a single expression has the probability of 1 and the others have the probability of 0, i.e., a deterministic string case; he selected the expression without hesitation. These expressions as well as the distributions should be highly organized and structured with complex semantic relationships.

Considering the above-mentioned observation, we give the following criteria for formulating the organized complexity measures.

1. The objects should be probability distributions. In addition, deterministic strings (as a special case of distributions) and more general distributions should be treated in a unified manner, e.g., the complexity of Cases 1, 2 and 3 for the source of Shakespeare's string should be measured in a unified manner.
2. Simple (or very regular) objects, which are treated as "problems of simplicity" based on Weaver's classification, should have low organized complexity.
3. Simply random objects, which are treated as "problems of disorganized complexity" by Weaver, e.g., the source of the chimpanzee's string, should have low organized complexity.
4. Highly organized objects, which are treated as "problems of organized complexity" by Weaver, e.g., Cases 1, 2 and 3 for the source of Shakespeare's string, should have high organized complexity.
5. The organized complexity of an object should be computable (or recursive in computation theory).

C. Existing Complexity Measures

Using these criteria, we now survey the typical quantitative definitions of organized complexity in literature.

Objects are “deterministic strings”

• Kolmogorov complexity

The notion of the Kolmogorov complexity was independently proposed by Solomonoff, Kolmogorov, and Chaitin [6, 22, 27, 37].

Roughly, the Kolmogorov complexity of string x is the size of the shortest program (on a computer) to produce string x .

More precisely, let U be a reference universal prefix (Turing) machine (see [27] for the reference universal prefix machine). Then, the Kolmogorov complexity, $K(x)$, of string $x \in \{0, 1\}^*$ is defined by

$$K(x) = \min\{|z| \mid U(z) = x, z \in \{0, 1\}^*\}.$$

In light of the above-mentioned criteria, the Kolmogorov complexity has the following properties.

1. The objects are only deterministic strings (Bad).
2. Simple (or very regular) objects have low Kolmogorov complexity (Good).
3. Simply random objects, deterministic n -bit strings, that are uniformly and randomly chosen from $\{0, 1\}^n$ have high Kolmogorov complexity (Bad).
4. Highly organized objects may have between high and low logical depth (Bad), since some highly organized complex objects that are generated from small strings through very long running-time and complex computations may have low Kolmogorov complexity. In other words, some organized complexities may be characterized in a dynamic manner as logical depth rather than in a static manner as Kolmogorov complexity.
5. The Kolmogorov complexity of an object (string) is not computable (Bad).

• Logical depth

Bennett [4] introduced *logical depth* with the intuition that complex objects are those whose most plausible explanations describe long causal processes. To formalize the intuition, Bennett employs the methodology of algorithmic information theory, the Kolmogorov complexity.

The logical depth of a deterministic string, Bennett’s definition for measuring organized complexity, is dependent on the running time of the programs that produce the string and whose length is relatively close to the minimum in a sense.

More precisely, the logical depth of string x at significance level $\epsilon := 2^{-b}$ [27] is

$$\min\{t \mid m_t(x)/m(x) \geq \epsilon\},$$

where we define $m_t(x)$ and $m(x)$ by

$$m_t(x) := \sum_{U^t(p)=x} 2^{-l(p)},$$

$$m(x) := \sum_{U(p)=x} 2^{-l(p)}.$$

Here, U is the reference universal prefix (Turing) machine (for the Kolmogorov complexity) and U^t is a specific class of U whose running time is bounded by t steps. $l(p)$ is the length of program p .

In light of the above-mentioned criteria, the logical depth has the following properties.

1. The objects are only deterministic strings (Bad).
2. Simple (or very regular) objects have low logical depth (Good).
3. Simply random objects have low logical depth (Good).
4. Highly organized objects may have between high and low logical depth (Bad), since some highly organized complex objects may have low logical depth with relatively high Kolmogorov complexity, where the core part of the organized complexity is due to the Kolmogorov complexity. In other words, some organized complexities may be characterized in a static manner as Kolmogorov complexity rather than in a dynamic manner as logical depth, (where $m_t(x)/m(x) \geq \epsilon$ is required but the logical depth is not dependent on the value of $m(x)$ itself (roughly, $-\log m(x)$ is close to the Kolmogorov complexity of x)).
5. The logical depth of an object (string) is not computable, since it is based on the Kolmogorov complexity or universal Turing machines (Bad).

• Effective complexity

Effective complexity [15, 16] was introduced by Gell-Mann, and is based on the Kolmogorov complexity. To define the complexity of an object, Gell-Mann considers the shortest description of the distribution in which the object is embedded as a typical member. Here, ‘typical’ means that the negative logarithm of its probability is approximately equal to the entropy of the distribution.

That is, the effective complexity of string x is

$$\min\{K(E) \mid -\log \Pr_E(x) \approx H(E)\},$$

where $K(E)$ is the Kolmogorov complexity of distribution E , i.e., the length of the shortest program to list all members, r , of E together with their probabilities, $\Pr_E(r)$, and $H(E)$ is the (Shannon) entropy of E .

In light of the above-mentioned criteria, the effective complexity has the following properties.

1. The objects are only deterministic strings (Bad). Technically, however, we can consider distribution E to be an object of the effective complexity.
2. Simple (or very regular) objects have low effective complexity (Good).
3. Simply random objects have low effective complexity (Good).
4. Highly organized objects may have between high and low effective complexity (Bad), since some highly organized complex objects may have low effective complexity with very high computational complexity of the universal machine to generate E , where the core part of the organized complexity is due to the computational complexity in a dynamic manner (e.g., high logical depth of E).
5. The effective complexity of an object (string) is not computable, since it is based on the Kolmogorov complexity or universal Turing machines (Bad).

Objects are “probability distributions”

• Thermodynamic depth

The *thermodynamic depth* was introduced by Lloyd and Pagels [28] and shares some informal motivation with logical depth, where complexity is considered a property of the evolution of an object.

We now assume the set of histories or trajectories that result in object (distribution) S_0 . A trajectory is an ordered set of macroscopic states (distributions) $S_{-L-1}, \dots, S_{-1}, S_0$. The thermodynamic depth of object S_0 is

$$H(S_{-L+1}, \dots, S_{-1} \mid S_0),$$

where $H(A, \dots, B \mid C)$ is the conditional entropy of combined distribution (A, \dots, B) with condition C .

One of the major problems with this notion is that it is not defined how long the trajectories (what value of L) should be. Moreover, it is impossible to specify formally the trajectories, given an object, since there is no description on how to select macroscopic states in [28]. If there are thousands of possible sets of macroscopic states, we would have thousands of different definitions of the thermodynamic depth.

Another fundamental problem with this measure is that in order to measure the complexity of an object S_0 , a set of macroscopic states $\mathcal{S} := \{S_{-L+1}, \dots, S_{-1}\}$ whose complexity is comparable to or more than that of S_0 should be established beforehand. Hence, if \mathcal{S} is fixed, or the thermodynamic depth of \mathcal{S} is concretely defined, it cannot measure the complexity of an object whose complexity is more than that of \mathcal{S} . That is, any concrete definition of this notion can measure only a restricted subset of objects, i.e., any concrete and generic definition is impossible in thermodynamic depth. It should be a fundamental problem with this concept.

As a result, it is difficult to define rigorously the thermodynamic depth and to characterize the definition.

Note that we have the same criticisms for the existing semantic information theories that are described in Section III A.

• Effective measure complexity

The *effective measure complexity* was introduced by Grassberger [18] and measures the average amount by which the uncertainty of a symbol in a string decreases due to the knowledge of previous symbols.

For distribution X^N over $\{0, 1\}^N$ ($N \in \mathbb{N}$), $H(X^N)$ is the Shannon entropy of X^N . Let $h_N := H(X^{N+1}) - H(X^N)$, and $h := \lim_{N \rightarrow \infty} h_N$. The effective measure complexity of $\{X^N\}_{N \in \mathbb{N}}$ is

$$\sum_{N=0}^{\infty} (h_N - h)$$

This difference quantifies the perceived randomness which, after further observation, is discovered to be order [25].

In light of the above-mentioned criteria, the effective measure complexity has the following properties:

1. The objects are only probability distributions, and deterministic strings are outside the scope of this measure (the complexity is 0 for any deterministic string) (Bad).
2. Simple (or very regular) objects have low effective measure complexity (Good).
3. Simply random objects have low effective measure complexity (Good).
4. Highly organized objects may have between high and low effective measure complexity (Bad), since some highly organized complex objects (distributions) may have low effective measure complexity with very high Kolmogorov complexity or computational complexity of the universal machine to generate

them, where the core of the organized complexity is due to some Kolmogorov complexity or the computational complexity. In other words, the effective measure complexity cares only about distributions but not the computational features that logical depth and effective complexity care about.

5. The effective measure complexity of an object (distribution) is computable, since it is not based on any Turing machine (Good).

• Statistical complexity

The *statistical complexity* was introduced by Crutchfield and Young [8]. Here, to define the complexity, the set of causal states S and the probabilistic transitions between them are modeled in the so-called ϵ -machine, which produces a stationary distribution of causal states, D_S . The mathematical structure of the ϵ -machine is a stochastic finite-state automaton or hidden Markov model.

Let S_i for $i = 1, \dots, k$ be causal states, $S := \{S_1, \dots, S_k\}$ and T_{ij} be the probability of a transition from state S_i to state S_j , i.e., $T_{ij} := \Pr[S_j | S_i]$. Each transition from S_i to S_j is associated with an output symbol, σ_{ij} , (e.g., $\sigma_{ij} \in \{0, 1\}$). Then, $\Pr[S_i]$, the probability that S_i occurs in the infinite run of the ϵ -machine, is given by the eigenvector of matrix $T := (T_{ij})$, since $\sum_{i=1}^k \Pr[S_i] \cdot T_{ij} = \Pr[S_j]$, i.e., $(\Pr[S_1], \dots, \Pr[S_k]) \cdot T = (\Pr[S_1], \dots, \Pr[S_k])$. Hence, the machine produces a stationary distribution of states, D_S . The output of the ϵ -machine is the infinite sequence of σ_{ij} induced by the infinite sequence of the transition of states. That is, ϵ -machine outputs a distribution, Σ_S , over $\{0, 1\}^\infty$, induced by D_S .

The statistical complexity of object X (distribution), denoted C_1 , is the minimum value of the Shannon entropy of D_S , $H_1(D_S)$, when $\Sigma_S = X$:

$$C_1 := \min\{H_1(D_S) \mid \Sigma_S = X\}.$$

A more generalized notion, C_α ($0 \leq \alpha \leq \infty$), is defined by the Reny entropy of D_S in place of the Shannon entropy, i.e.,

$$C_\alpha := \min\{H_\alpha(D_S) \mid \Sigma_S = X\},$$

where C_1 is the case where $\alpha := 1$ as H_1 is the Shannon entropy, and $C_0 := \min\{\log \#S \mid \Sigma_S = X\}$ ($\alpha := 0$) ($\#S$ is the number of elements of set S) (For $\alpha := \infty$, H_∞ is the mini-entropy).

In light of the above-mentioned criteria, the statistical complexity has the following properties:

1. Statistical complexity C_1 can measure only probability distributions as objects, since

$C_1 = 0$ for any deterministic string. Complexity C_0 cannot measure the deterministic strings well either, since the organized complexity may be around $|S|$ for high Kolmogorov complexity or high logical depth deterministic strings but $C_0 = \log |S|$. Moreover, C_0 cannot capture the distribution of the ϵ -machine, since it only depends on the number of vertexes of causal states. That is, none of C_α with a value of α ($0 \leq \alpha \leq \infty$) can measure probability distributions and deterministic strings in a unified manner (Bad).

2. Simple (or very regular) objects have low statistical complexity (Good).
3. Simply random objects have low statistical complexity (Good).
4. As for the standard definition of statistical complexity, i.e., $\alpha := 1$ or the Shannon entropy, highly organized objects may have between high and low statistical complexity (Bad), since (1) some highly organized complex objects (almost deterministic strings) may have low statistical complexity, almost zero, where the core of the organized complexity is due to the complexity of the almost deterministic data part, and (2) some highly organized complex objects (distributions) have relatively low statistical complexity with highly complex output mapping $\{\sigma_{ij}\}$ of ϵ -machines, where the core of the organized complexity is due to the complexity of $\{\sigma_{ij}\}$ of ϵ -machines (statistical complexity C_α depends on only D_S but is independent of the complexity of output mapping $\{\sigma_{ij}\}$). See Remark 4 for more precise observation.
5. The statistical complexity of an object (distribution/strings) is computable, since it is not based on any Turing machines (Good).

In summary, the existing measures have the following drawbacks.

- Every existing complexity measure focuses on a single feature of complexity, for example, logical depth focuses on the computational complexity feature, the Kolmogorov complexity and effective complexity focus on the descriptive complexity feature, and statistical complexity focuses on the distributional complexity feature, but no existing measure captures all of them simultaneously.
- Every existing complexity measure can treat either a probabilistic or deterministic form of an object, but no measure can cover both in a unified manner.
- Some of the measures, the Kolmogorov complexity, effective complexity and logical depth, that are based on Turing machines are not computable.

D. Proposed Organized Complexity Measure

We now propose a new quantitative definition of organized complexity. Roughly speaking, the proposed quantitative definition is given by the shortest description size of a (stochastic finite-state automaton form of) *circuit* [34, 39] that simulates an object (probability distribution).

In the existing complexity measures surveyed in Section IIC, some computing machineries are employed. In the logical depth and effective complexity, universal Turing machines are employed, which cause the uncomputability of their measures. In the effective measure complexity, no computational machinery is used; hence, it cannot capture the computational and descriptonal features of organized complexities, which logical depth and effective complexity capture, respectively. The statistical complexity employs ϵ -machines, whose mathematical model is a stochastic finite-state automaton or hidden Markov model; hence it captures the distributional features of organized complexities but not the computational, descriptonal, and deterministic-object features. See Remark 4 for more details.

In the place of universal Turing machines and ϵ -machines, we employ another class of machinery, a stochastic finite-state automaton form of *circuit*, *oc-circuit*. Our new measure is given by the *shortest* description size of an oc-circuit for simulating the object. That is, Occam's razor plays a key role in our definition. The advantage of using circuits is that it can capture the computational and distributional features of complexity as the size of a circuit as well as the descriptonal features of complexity as the input size of a circuit. Moreover, the shortest description size of an oc-circuit for simulating an object is computable (Theorem 1), in contrast to that in which the shortest program size on a Turing machine is uncomputable [27]. Our approach is more general than the approach by ϵ -machines in the statistical complexity, since our oc-circuit model can simulate any ϵ -machine as a special case (Theorem 2).

The major difference between circuits and Turing machines is that a single (universal) Turing machine can compute any size of input, while a single circuit can compute a fixed size of input. In spite of the difference, any bounded time computation of a Turing machine can be computed by a bounded size of a circuit [34, 39]. Hence, the proposed complexity measure based on circuits captures general computational features in complexity. In addition, the finiteness of each circuit yields the computability of our measure, in contrast to the uncomputability of Turing machine based measures such as logical depth and effective complexity as well as the Kolmogorov complexity.

We now define a new measure of organized complexity. First, we define our computation model, *oc-circuit*.

Definition 1 (OC-Circuit) Let circuit C with N input bits and L output bits be a directed acyclic graph in which

every vertex is either an input gate of in-degree 0 labeled by one of the N input bits, or one of the basis of gates $B := \{AND, OR, NOT\}$. Among them, L gates are designated as the output gates. That is, circuit C actualizes a Boolean function: $\{0, 1\}^N \rightarrow \{0, 1\}^L$.

Let $s_i \in \{0, 1\}^{N_s}$ be a state at step i ($i \in \mathbb{N}$), $u \in \{0, 1\}^{N_u}$ be an a priori input (universe), $m_i \in \{0, 1\}^{N_m}$ be an input at step i , $r_i \stackrel{U}{\leftarrow} \{0, 1\}^{N_r}$ be random bits at step i , and $N := N_u + N_s + N_m + N_r$. Then,

$$(s_{i+1}, y_i) \leftarrow \boxed{C(u, \cdot)} \leftarrow (s_i, m_i, r_i), \quad i = 1, 2, \dots, K,$$

i.e., $(s_{i+1}, y_i) := C(u, s_i, m_i, r_i)$, where $y_i \in \{0, 1\}^{L_y}$, $N_m \leq L_y$ is the output of C at step i , and $L := N_s + L_y$. Let V be the number of vertexes of C .

Let $\tilde{C} := ((w_{ij})_{i=1, \dots, V; j=1, \dots, V}, (\ell_1, \dots, \ell_V), (o_1, \dots, o_L))$ be a canonical description of C , where $(w_{ij})_{i=1, \dots, V; j=1, \dots, V}$ is the adjacent matrix of directed graph C , i.e., $w_{ij} := 1$ iff there is an edge from vertex i to vertex j , and $w_{ij} := 0$ otherwise, ℓ_i ($i = 1, \dots, V$) is the label of the i -th vertex, $\ell_i \in \{1, \dots, N, AND, OR, NOT\}$, i.e., each vertex i is labeled by ℓ_i , and $o_i \in \{1, \dots, V\}$ is the vertex designated to the i 's output, i.e., (o_1, \dots, o_L) is the sequence of output gates. Hereafter, we abuse the notation of C to denote \tilde{C} , the canonical description of C .

Let $\mathcal{C} := (\overline{C}, u, n, \vec{m})$ be an "oc-circuit", and Y be the output of \mathcal{C} , where $\overline{C} := (C, N_u, N_s, N_m, N_r, L_y, s_1)$, $K := \lceil n/L_y \rceil$, $\vec{m} := (m_1, \dots, m_K)$, $Y := (y_1, \dots, y_K)_n$ (see Section IC for the notation of $(\dots)_n$).

The output, Y , of \mathcal{C} can be expressed by $Y \stackrel{R}{\leftarrow} \mathcal{C}$, i.e., $Y \stackrel{R}{\leftarrow} (\overline{C}, u, n, \vec{m}_n)$, where the probability of distribution Y is taken over the randomness of $r_i \stackrel{U}{\leftarrow} \{0, 1\}^{N_r}$ ($i = 1, \dots, K$).

Then, \overline{C} , u , and \vec{m} are called the "logic," "universe," and "semantics" of oc-circuit \mathcal{C} , respectively.

Remark 1 Circuit C of oc-circuit \mathcal{C} is a probabilistic circuit, where uniformly random strings, $r_i \stackrel{U}{\leftarrow} \{0, 1\}^{N_r}$ for $i = 1, \dots, K$, are input to C and the output of C is distributed over the random space of $\{r_i\}_{i=1, \dots, K}$.

Here note that $\{r_i\}_{i=1, \dots, K}$, which is an input to C , is not included in \mathcal{C} , while the other inputs to C , u and $\{m_i\}_{i=1, \dots, K}$, are included in \mathcal{C} . In other words, the size of the randomness, $\sum_{i=1}^K |r_i|$, is ignored in the size of \mathcal{C} or the definition of the organized complexity (see Definition 2), while N_r and a part of C regarding the randomness are included in \mathcal{C} . This is because the randomness, $\{r_i\}_{i=1, \dots, K}$, is just the random source of \mathcal{C} 's output distribution and has no organized complexity itself. Hence, simply random objects are characterized to have low organized complexity based on the size of oc-circuit \mathcal{C} (see item 3 in the property summary of the proposed complexity measure in the end of this section).

Remark 2 Although the parts of an oc-circuit, \overline{C} , u , and \vec{m} , are named logic, universe, and semantics, respectively, we do not care about the meanings of these names

in Section IID. We care more about these meanings in Section IIIB.

Definition 2 (Organized Complexity)

Let X be a distribution over $\{0,1\}^n$ for some $n \in \mathbb{N}$.

“Organized complexity” OC of distribution X at precision level δ ($0 \leq \delta < 1$) is

$$\text{OC}(X, \delta) := \min\{|\mathcal{C}| \mid X \overset{\delta}{\approx} Y \overset{\text{R}}{\leftarrow} \mathcal{C}\}, \quad (1)$$

where $\mathcal{C} := (\overline{\mathcal{C}}, u, n, \vec{m})$ is an oc-circuit, and $|\mathcal{C}|$ denotes the bit length of the binary expression of \mathcal{C} (see Section IC for the notations of $\overset{\delta}{\approx}$).

We call oc-circuit $\mathcal{C}^X := (\overline{\mathcal{C}}^X, u^X, n, \vec{m}^X)$ the shortest (or proper) oc-circuit of X at precision level δ , if $X \overset{\delta}{\approx} Y \overset{\text{R}}{\leftarrow} \mathcal{C}^X$ and $|\mathcal{C}^X| = \text{OC}(X, \delta)$. If there are multiple shortest oc-circuits of X , i.e., they have the same bit length, the lexicographically first shortest one is selected as the shortest oc-circuit.

Then, $\overline{\mathcal{C}}^X$, u^X , and \vec{m}^X are called the “proper logic,” “proper universe,” and “proper semantics” of X at precision level δ , respectively. Here, $X \overset{\delta}{\approx} Y \overset{\text{R}}{\leftarrow} \mathcal{C}^X := (\overline{\mathcal{C}}^X, u^X, n, \vec{m}^X)$.

Theorem 1 For any distribution X over $\{0,1\}^n$ ($n \in \mathbb{N}$) and any precision level $\delta > 0$, $\text{OC}(X, \delta)$ can be computed.

Proof

For any distribution $X := \{(x, p_x) \mid x \in \{0,1\}^n, p_x \in [0,1], \sum_{x \in \{0,1\}^n} p_x = 1\}$ and any precision level $\delta > 0$, there always exists another distribution $X' := \{(x, p'_x) \mid x \in \{0,1\}^n, 0 \leq p'_x \leq 1, p'_x \in \mathbb{Q}, \sum_{x \in \{0,1\}^n} p'_x = 1\}$ such that $X' \overset{\delta}{\approx} X$ (Here note that $p_x \in \mathbb{R}$ is changed to $p'_x \in \mathbb{Q}$ provided that $X' \overset{\delta}{\approx} X$).

We then construct the truth table of Boolean function $f : \{0,1\}^\ell \rightarrow \{0,1\}^n$ such that $\Pr[x = f(r) \mid r \overset{\text{U}}{\leftarrow} \{0,1\}^\ell] = p'_x$ for all $x \in \{0,1\}^n$, where the probability is taken over $\overset{\text{U}}{\leftarrow} \{0,1\}^\ell$. Such a function, f , can be achieved by setting truth table $T_f := \{(r, f(r))\}_{r \in \{0,1\}^\ell}$ such that $\#\{r \mid f(r) = x\}/2^\ell = p'_x \in \mathbb{Q}$ for all $x \in \{0,1\}^n$.

Since any Boolean function can be achieved by a circuit with basis $B := \{\text{AND}, \text{OR}, \text{NOT}\}$ [12], we construct circuit \mathcal{C}^* for oc-circuit \mathcal{C}^* with $N_s := 1, N_u = N_m := 0$ (i.e., $u = m_i := \lambda$), $N_r := \ell, K := 1, L_y := n, s_1 = s_2 := 0$. That is, $(0, y_1) := \mathcal{C}^*(\lambda, 0, \lambda, r_1)$, and the output of \mathcal{C}^* is $y_1 \in \{0,1\}^n$ with the same distribution as that of X' over the randomness of $r_1 \overset{\text{U}}{\leftarrow} \{0,1\}^{N_r}$. That is, $y_1 \overset{\text{R}}{\leftarrow} \mathcal{C}^*$ and $X \overset{\delta}{\approx} X' = y_1$.

From the definition of OC , $\text{OC}(X, \delta) \leq |\mathcal{C}^*|$.

We then, exhaustively check all values of Z with $|Z| < |\mathcal{C}^*|$ whether Z is an oc-circuit such that $X \overset{\delta}{\approx} Y \overset{\text{R}}{\leftarrow} Z$. Here note that we can syntactically check whether or not Z is the correct form of an oc-circuit. Finally, we find the shortest one among the collection of Z (and \mathcal{C}^*)

satisfying the condition. Clearly, the size of the shortest one is $\text{OC}(X, \delta)$. \square

Remark 3 As clarified in this proof, given object (distribution) X and precision level δ , the proposed definition of organized complexity uniquely determines (computes) not only organized complexity $\text{OC}(X, \delta)$ but also the shortest (proper) oc-circuit, \mathcal{C}^X , including proper logic $\overline{\mathcal{C}}^X$, proper universe u^X , and proper semantics \vec{m}^X of X . In other words, the definition characterizes the complexity features of object X , i.e., it characterizes not only organized complexity $\text{OC}(X, \delta)$, but also structural complexity features of X , e.g., computational and distributional features by the size of $\overline{\mathcal{C}}^X$ and descriptive features by the size of u^X and \vec{m}^X .

In the following theorem, we show that the notion of oc-circuit with the proposed organized complexity includes the ϵ -machine with statistical complexity introduced in Section IIC as a special case.

Theorem 2 Any ϵ -machine can be simulated by an oc-circuit.

Proof

Given ϵ -machine, $(\{S_1, \dots, S_k\}, (T_{ij})_{i=1, \dots, k; j=1, \dots, k}, (\sigma_{ij})_{i=1, \dots, k; j=1, \dots, k})$, we construct oc-circuit $\mathcal{C} := ((\mathcal{C}, N_u, N_s, N_m, N_r, L_y, s_1), u, n, (m_1, m_2, \dots))$ such that $N_s := \lceil \log_2 k \rceil + 1$ (i.e., $S_i \in \{0,1\}^{N_s}$), $N_u := 0$, $N_m := 0$, $N_r := \max_{i,j} \{|T_{ij}|\}$, $L_y := \max_{i,j} \{|\sigma_{ij}|\}$, $s_1 := S_1$ (initial causal state), $n := \infty$, $u := \lambda$ (null string), $m_i := \lambda$ ($i = 1, 2, \dots$) and \mathcal{C} is achieved to satisfy $\Pr[(S_j, \sigma_{ij}) := \mathcal{C}(\lambda, S_i, \lambda, r)] = T_{ij}$ for $i, j = 1, \dots, k$, where $|T_{ij}|$ is the bit length of the binary expression of T_{ij} , and the probability is taken over the randomness of $r \overset{\text{U}}{\leftarrow} \{0,1\}^{N_r}$ in each execution of \mathcal{C} . It is clear that the behavior of this oc-circuit with respect to the causal states is exactly the same as that for the given ϵ -machine. \square

Remark 4 (Features of the proposed complexity) The proposed organized complexity is characterized by the minimum length of the description of whole oc-circuit \mathcal{C} , but the statistical complexity is characterized by some partial information on \mathcal{C} , i.e., only the average size of a compressed coding of a causal state, $H(D_S) \leq N_s$, for $\alpha = 1$ (\mathcal{C}_1), or the (uncompressed) size of a causal state, N_s , for $\alpha = 0$ (\mathcal{C}_0). That is, our complexity measure captures the complexity of whole circuit (logic) $\overline{\mathcal{C}}$, while the statistical complexity only captures a partial property of the “distributional complexity” of $\overline{\mathcal{C}}$, the compressed or uncompressed size of a causal state, but ignores the distributional complexity of T_{ij} and σ_{ij} (expressed by N_r and the complexity of $\overline{\mathcal{C}}$). That is, even if we only focus on the distributional complexity features (where $N_u = N_m = 0$), the statistical complexity only captures some of the features, while our definition captures the whole as the size of $\overline{\mathcal{C}}$ including N_s and N_r .

In addition, our complexity measure can treat more general cases with $N_u > 0$ and $N_m > 0$, while statistical complexity only considers a limited case with $N_u = 0$ and $N_m = 0$, i.e., it ignores the descriptonal complexity features as well as the computational features. For example, a sequence in a genome pattern that is common to all individuals is considered to be determined in the evolution process, and has some biological meaning. The biological knowledge of DNA necessary to understand the DNA sequences can be captured by logic \overline{C} and universe u of the oc-circuit C (where $|u| = N_u > 0$), and the characteristic information (biological meaning) on each genome pattern ($\overset{\delta}{\approx} Y \overset{R}{\leftarrow} C$) can be captured by semantics \vec{m} of C (where $|m_i| = N_m > 0$).

Moreover, our complexity definition covers more complexity features. The “computational complexity” features of an object, which are captured by the logical depth, are characterized by the size of logic \overline{C} of oc-circuit C in our definition, and the “descriptonal complexity” features of an object, which are captured by the effective complexity (and Kolmogorov complexity), are characterized by the size of universe u and semantics \vec{m}_n of the oc-circuit C .

We can achieve a circuit C using another basis of gates, e.g., $\{NAND\}$ and $\{AND, NOT\}$, in place of $\{AND, OR, NOT\}$. We express an oc-circuit using such a basis by $C_{\{NAND\}}$ and $C_{\{AND, NOT\}}$, respectively. We also express the organized complexity of X using such a circuit by $OC_{\{NAND\}}(X, \delta)$ and $OC_{\{AND, NOT\}}(X, \delta)$, respectively.

Based on such a different basis of gates, a natural variant of the proposed organized complexity, *structured organized complexity*, is given below.

Definition 3 (Structured Organized Complexity)

Let X be a distribution over $\{0, 1\}^n$ for some $n \in \mathbb{N}$.

Let $C^S := (\overline{C}^S, u, n, \vec{m})$ be a structured oc-circuit, where $\overline{C}^S = (\text{macros}, C_{\text{macros}}, N_u, N_s, N_m, N_r, L_y, s_1)$ and macros represents a set of macro gates (subroutine circuits) that are constructed from basis gates and that can be hierarchically constructed, where a level of macro gates are constructed from lower levels of macro gates. In addition, macros is notationally abused as the canonical description of macro gates in macros , which is specified in the same manner as that in the canonical description of a circuit. Term C_{macros} is (the canonical description of) a circuit constructed from basis gates B as well as macro gates in macros .

Structured organized complexity OC^S of distribution X at precision level δ ($0 \leq \delta < 1$) is

$$OC^S(X, \delta) := \min\{|\mathcal{C}^S| \mid X \overset{\delta}{\approx} Y \overset{R}{\leftarrow} \mathcal{C}^S\}.$$

Theorem 3 For any distribution X over $\{0, 1\}^n$ ($n \in \mathbb{N}$) and any precision level $\delta > 0$, $OC^S(X, \delta)$ can be computed.

Proof Given distribution X , we can construct structured oc-circuit C^* in the same manner as that shown

in the proof of Theorem 1. Here note that any (basic) oc-circuit C can be expressed as structured oc-circuit C^S where $\text{macros} := \lambda$, with slightly relaxing the format for structured oc-circuits, or to allow $\text{macros} := \lambda$.

From the definition of OC^S , for any value of $0 < \delta < 1$, $OC^S(X, \delta) \leq |C^*|$.

We then, given δ , exhaustively check all values of Z with $|Z| < |C^*|$ whether Z is a structured oc-circuit such that $X \overset{\delta}{\approx} Y \overset{R}{\leftarrow} Z$. Finally, we select the shortest one among the collection of Z (and C^*) satisfying the condition. Clearly, the size of the shortest one is $OC^S(X, \delta)$. \square

Remark 5 As described in Remark 3, the shortest structured oc-circuit with these parameters characterizes the properties of object X . It especially shows the optimized hierarchically structured circuit C^S .

Remark 6 As mentioned above, we can construct a structured oc-circuit using another basis such as $\{NAND\}$ and $\{AND, NOT\}$, e.g., $C_{\{NAND\}}^S$ and $C_{\{AND, NOT\}}^S$. Then, macros in $C_{\{NAND\}}^S$ can consist of macro gates of AND , OR , and NOT from $NAND$ gates. Since the size of macros is a constant $O(1)$ in n ,

$$OC^S(X, \delta) \leq OC_{\{NAND\}}^S(X, \delta) \leq OC^S(X, \delta) + O(1).$$

Variations:

We have more variations of the organized complexity.

- (Computational distance) In Definitions 2 and 3, statistical distance is used for defining the closeness $\overset{\delta}{\approx}$. We can replace this with the “computational” closeness $\overset{(D, \delta)}{\approx}$. Here, for two distributions, X and Y , over $\{0, 1\}^n$, the computational closeness of X and Y is defined by
$$X \overset{(D, \delta)}{\approx} Y \quad \text{iff} \quad \forall D \in \mathcal{D}, \quad \frac{1}{2} \cdot |\Pr[1 \overset{R}{\leftarrow} D(\alpha) \mid \alpha \overset{R}{\leftarrow} X] - \Pr[1 \overset{R}{\leftarrow} D(\alpha) \mid \alpha \overset{R}{\leftarrow} Y]| < \delta,$$
where \mathcal{D} is a class of machines $D : \{0, 1\}^n \rightarrow \{0, 1\}$. Intuitively, the computational closeness means that X and Y are indistinguishable at precision level δ by any machine in class \mathcal{D} .
- (Quantum circuits) Circuit C in Definitions 2 and 3 can be replaced by a “quantum” circuit [30]. In this variation, we assume that the source of a distribution (observed data) is principally given by a quantum phenomenon.

There are several variations of the quantum complexity definition, typically: (1) all inputs and outputs of C are classical strings, (2) only state s_i is a quantum string and the others are classical, and (3) all inputs and outputs of C except output y_i are quantum strings.

3. (Interactions) If an observation object actively reacts similar to a living thing, we often observe it in an interactive manner.

So far in this paper we have assumed that an object is a distribution that we perceive passively. We can extend the object from such a passive one to an active one with interactive observation.

Suppose that the observation process is interactive between observer A and observation object B . For example, A first sends z_1 to B , which replies x_1 to A , and we continue the interactive process, $z_2, x_2, \dots, z_J, x_J$.

Let $Z := (z_1, \dots, z_J)$ and $X := (x_1, \dots, x_J)$ be distributions. We then define the conditional organized complexity of X under Z with precision level δ , $\text{OC}(X : Z, \delta)$, which can be defined as the shortest (finite version of) conditional oc-circuit to simulate X (with precision level δ) under Z (see Definition 8 for the conditional oc-circuit).

The extended notion of the organized complexity of interactive object (X, Z) can be defined by $\text{OC}(X : Z, \delta)$.

In light of the criteria described in Section II B, the proposed complexity measure has the following properties.

1. The proposed complexity definition covers probability distributions and deterministic strings (as special cases of distributions) in a unified manner (Good).

For example, for any deterministic string (as a special case of distribution), there exists an oc-circuit with circuit C to simulate the deterministic string by $Y := (y_1, \dots, y_K)_n \in \{0, 1\}^n$, such that

$$(s_{i+1}, y_i) := C(u, s_i, m_i, \lambda), \quad i = 1, 2, \dots, K,$$

where $N_r = 0$, i.e., $r_i := \lambda$ for $i = 1, \dots, K$.

2. Simple (or very regular) objects have low complexity (Good).

For example, in a very regular case ('11...1' $\in \{0, 1\}^n$), the object can be simulated by an oc-circuit $\mathcal{C} := (\overline{C}, 1, n, \lambda)$ with $\overline{C} := (C, 1, 1, 0, 0, 1, 0)$, such that

$$(s_{i+1}, 1) := C(1, s_i, \lambda, \lambda), \quad i = 1, 2, \dots, n,$$

where $N_u = 1 (u = 1)$, $s_i := 0$ for $i = 1, \dots, n + 1$ (i.e., $N_s = 1$), $N_m = N_r = 0$ (i.e., $m_i = r_i = \lambda$) and $L_y = 1$. That is, input size $N = 2$ and output size $L = 2$, i.e., C has 2 gates that are input gates labeled by $(1, 2)$ and that are also output gates. $C := ((w_{ij}) := I_2, (\ell_1, \ell_2) := (1, 2), (o_1, o_2) := (2, 1))$, where I_2 is the 2-dimensional identity matrix. Hence, $\mathcal{C} :=$

$(((((1, 0, 0, 1), (1, 2), (2, 1)), 1, 1, 0, 0, 1, 0), 1, n, \lambda)$, and $\text{OC}('11...1') \leq c + \log_2 n$, where c is a small constant.

3. Simply random objects have low complexity (Good).

For example, in the case of uniformly random distribution X over $\{0, 1\}^n$, the object can be simulated by an oc circuit $\mathcal{C} := (\overline{C}, \lambda, n, \lambda)$ with $\overline{C} := (C, 0, 1, 0, 1, 1, 0)$, such that

$$(s_{i+1}, r_i) := C(\lambda, s_i, \lambda, r_i), \quad i = 1, 2, \dots, n,$$

where $N_u = 0$, $s_i := 0$ for $i = 1, \dots, n$ (i.e., $N_s = 1$), $N_m = 0$ (i.e., $u = m_i = \lambda$), $N_r = 1$, $r_i \stackrel{U}{\leftarrow} \{0, 1\}$, and $L_y = 1$. That is, input size $N = 2$ and output size $L = 2$, i.e., C has 2 gates that are input gates labeled by $(1, 2)$ and that are also output gates. $C := ((w_{ij}) := I_2, (\ell_1, \ell_2) := (1, 2), (o_1, o_2) := (1, 2))$. Hence, $\mathcal{C} := (((((1, 0, 0, 1), (1, 2), (1, 2)), 0, 1, 0, 1, 1, 0), \lambda, n, \lambda)$, and $\text{OC}(X) \leq c + \log_2 n$, where c is a small constant.

4. Highly organized objects have high complexity (Good).

As described in Remarks 3 and 4, the proposed complexity definition simultaneously captures the distributional features of complexity (similar to statistical complexity and effective measure complexity), computational features of complexity (similar to logical depth), and descriptive features of complexity (similar to the Kolmogorov complexity and effective complexity).

Hence, our complexity measure does not have the drawbacks of the existing complexity measures described in Section II C, i.e., our measure correctly evaluates the complexity of highly organized objects for which the existing measures miss-evaluate to be low. In addition, objects evaluated by any existing (organized) complexity measure to be high for some features of complexity are also evaluated to be high in our complexity measure.

5. The proposed complexity of any object (distributions/strings) is computable as shown in Theorems 1 and 3 (Good).

III. SEMANTIC INFORMATION THEORY

A. Existing Theories and Problems

As described in Section I A, the semantic information theory has been studied for over six decades, e.g., [2, 3, 5, 11, 13, 14, 19–21, 23, 29, 31, 38, 42]. Among these studies, we here investigate the existing semantic

information theories that offer a quantitative measure of semantic information, e.g., [2, 3, 5, 11, 13, 14, 29, 31, 38].

Although they present many different ideas and approaches, a common paradigm among these theories assumes some a priori information, e.g., world model, knowledge database, and logic, to measure the amount of semantic information of an object.

For example, in [2], the following information is assumed to be established beforehand to define the semantic information measure.

- A world model that is a set of interpretations for propositional logic sentences with probability distributions.
- An inference procedure for propositional logic.
- A message generator that generates messages using some coding strategy.

It looks natural and inevitable to assume such a priori basic information such as the world model and logic to define the semantic information measure, but we have the same criticism for this paradigm as that for the thermodynamic depth [28] described in Section II C. That is, any existing semantic information theory in literature assumes such a priori information but gives no concrete or precise specification of the assumed a priori information. Without any concrete specification of the assumed information, we cannot rigorously define the semantic information amount and such a quantitative definition is just a vague and obscure notion. If there are thousands of possible concrete specifications of the information, we would have thousands of possible quantitative definitions.

In addition, when we try to measure the semantic information amount of an object, or we have no idea of its amount, such a priori information should be established beforehand and its complexity (information amount) should be comparable to or exceed that of the object. Hence, if the a priori information is fixed, or the semantic information measure with this information is concretely defined, it cannot measure the information amount of an object that has greater information amount than that of the a priori information. That is, any concrete definition in this paradigm can measure only a specific subset of objects, i.e., any concrete and generic definition is impossible, or any concrete definition is ad hoc. It should be an essential problem in the existing semantic information theories.

Another criticism of the existing semantic information theories is that they are constructed only on some mathematical logic such as propositional and first order logics. Our daily communications should be based on more complicated and fuzzy logic. It is well known that bees inform other bees of the direction and distance of flowers using their actions similar to dancing. Clearly some semantic information is transferred among bees in this case, and the logic for the semantics should be much different from mathematical logic and the logic of humans.

Given an object, e.g., Shakespeare's plays and bee's actions, we may roughly imagine which classes of information (universe) and logic are necessary or sufficient to understand the object. For example, in order to under-

stand the semantic information of Shakespeare's plays, the necessary universe and logic may be the knowledge of English sentences, the culture of that age and human daily logic. To understand the semantic information of bee communications, a much more limited and specific type of universe and logic may be sufficient.

We now have the following questions.

1. Can we quantitatively define the amount of semantic information of an object without assuming any a priori information? Or, can we quantitatively define the amount of semantic information of an object absolutely (not relative to a priori information)?
2. Given an object, can we determine the minimum amount of universe and logic to understand the semantic information of the object?

B. Proposed Semantic Information Theory

In this section, we present a mathematical theory of *semantic* information and communication that covers the semantic and effectiveness problems (Levels B and C of information and communication problems given by Weaver, see Section I A). The proposed theory is based on our organized complexity measure shown in Section II D. The proposed semantic information theory offers a positive answer to the questions raised at the end of Section III A.

We first consider an example described in Section II B, a source (distribution) from Shakespeare's plays.

Let distribution X over $\{0, 1\}^n$ be the source of several plays that consists of several hundred pages of English sentences, and let \mathcal{C}^X be the shortest (proper) oc-circuit, $(\overline{\mathcal{C}}^X, u^X, n, \vec{m}^X)$ to simulate source X at precision level δ . The output, $Y \stackrel{R}{\leftarrow} (\overline{\mathcal{C}}^X, u^X, n, \vec{m}^X)$, is the distribution statistically δ -close to X , the distribution of the source of the whole sentences in the plays.

The shortest oc-circuit, \mathcal{C}^X , for X can characterize source X such that

- the proper logic, $\overline{\mathcal{C}}^X$ of X may capture the features of Shakespeare's way of thinking and daily logic including English grammar,
- the proper universe of X , u^X , may capture the knowledge of English words and expressions as well as aspects of the cultures necessary to understand the plays,
- the proper semantics of X , \vec{m}^X , may capture the semantics (meanings) of sentences of the plays.

Based on the observation above, we formalize the semantic information theory.

In Section II D, we aim to define quantitatively the organized complexity of *physical* objects, i.e., the sources of physically observed data. Since any physical thing

is essentially bounded finitely, we assume a source is a distribution over finite strings, $\{0, 1\}^n$ ($n \in \mathbb{N}$), in Section IID.

In contrast, in this section, we aim to construct a *mathematical* theory of semantic information, where we focus on the *asymptotic* properties of an object when the size of the object is supposed to be increasing unlimitedly. This is because this theory focuses on the semantics part of the organized complexity, which consists of logic, universe, and semantics. The semantics part such as the proper semantics, \vec{m}^X , of object (source) X in the example above can be characterized more clearly and simply using the asymptotic properties (where the sizes of \vec{m}^X and X are supposed to be increasing unlimitedly while the sizes of proper logic \vec{C}^X and proper universe u^X are finitely bounded) than by the finite-size properties (where the size of \vec{m}^X is finitely bounded). Note that Shannon's information theory is also a theory for asymptotic properties. Therefore, an object here is not a single distribution but a *family* of distributions, $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$, where X_n is a distribution over $\{0, 1\}^n$.

We define several notions including the *semantic information amount*. Note that Occam's razor also plays a key role in this definition, since it is based on organized complexity, OC.

1. Semantic Information Amount

First we introduce the notion of a *family of distributions* and (naturally) extend the concept of an oc-circuit (for a distribution on finite-size strings) into that of an *oc-circuit for a family of distributions*, which is the same as the original one except that its output and semantics are unbounded in this concept, while they are bounded in the original.

Definition 4 (*Family of Distributions and OC-Circuit for a Family of Distributions*)

Let X_n be a distribution such that $X_n := \{(x, p_x) \mid x \in \{0, 1\}^n, 0 \leq p_x \leq 1, \sum_{x \in \{0, 1\}^n} p_x = 1\}$, and $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ be a “family of distributions.”

Let $\mathcal{C} := (\vec{C}, u, \infty, \vec{m}_\infty)$ be an “oc-circuit for a family of distributions” such that

$$\begin{aligned} \vec{C} &:= (C, N_u, N_s, N_m, N_r, L_y, s_1), \\ \vec{m}_\infty &:= (m_i)_{i=1,2,\dots} := (m_1, m_2, \dots) \in \{0, 1\}^\infty, \\ (s_{i+1}, y_i) &\leftarrow \boxed{C(u, \cdot)} \leftarrow (s_i, m_i, r_i), \quad i = 1, 2, \dots, \\ \text{i.e., } (s_{i+1}, y_i) &:= C(u, s_i, m_i, r_i), \quad i = 1, 2, \dots, \\ Y_n &:= (y_1, \dots, y_{K_n})_n, \quad K_n := \lceil n/L_y \rceil, \quad \text{for } n \in \mathbb{N}, \\ \mathcal{Y} &:= \{Y_n\}_{n \in \mathbb{N}} \stackrel{R}{\leftarrow} \mathcal{C}, \end{aligned}$$

where $N_m \leq L_y$, $m_i \in \{0, 1\}^{N_m}$, $u \in \{0, 1\}^{N_u}$, $s_i \in \{0, 1\}^{N_s}$, $r_i \in \{0, 1\}^{N_r}$, $y_i \in \{0, 1\}^{L_y}$ and $\{0, 1\}^\infty$ is the set of infinite-size binary strings.

Let $\mathcal{C}_n := (\vec{C}, u, n, \vec{m}_n)$, where $\vec{m}_n := (m_1, \dots, m_{K_n})$.

Definition 5 (*Sequential Family of Distributions*)

A family of distributions, $\mathcal{Y} := \{Y_n\}_{n \in \mathbb{N}}$, is called a “sequential family of distributions” if for any n' and n in \mathbb{N} with $n' > n$, distribution Y_n over $\{0, 1\}^n$ is the n -bit restriction of distribution $Y_{n'}$ over $\{0, 1\}^{n'}$ (see Section IC for the definition of n -bit restriction).

Remark 7 The family of distributions output by an oc-circuit for a family of distributions, $\mathcal{Y} := \{Y_n\}_{n \in \mathbb{N}}$, is a sequential family of distributions.

A value of $\mu := (\mu_1, \mu_2, \dots) \in \{0, 1\}^\infty$ ($\mu_i \in \{0, 1\}$ for $i = 1, 2, \dots$) corresponds to a value in $[0, 1] \subset \mathbb{R}$ by map $\varphi : \{0, 1\}^\infty \mapsto [0, 1]$, $\varphi : (\mu_1, \mu_2, \dots) \rightarrow “0.\mu_1\mu_2\dots” \in [0, 1]$, where “ $0.\mu_1\mu_2\dots$ ” is the binary expression of a value in $[0, 1]$.

Through this correspondence, if $\mathcal{Y} := \{Y_n\}_{n \in \mathbb{N}}$ is a sequential family of distributions, $\lim_{n \rightarrow \infty} Y_n$ corresponds to probability density function $Y(\cdot)$ with support $[0, 1]$

[7] such that $\int_0^1 Y(x)dx = 1$, and $Y_n = \{(x, p_x) \mid x \in \{0, 1\}^n, p_x := \int_{“0.x”}^{“0.x”+1/2^n} Y(x)dx\}$, where “ $0.x$ ” denotes “ $0.x_1x_2\dots x_n$ ” $\in [0, 1]$, if $x = (x_1, \dots, x_n) \in \{0, 1\}^n$.

Since there are a variety of unnatural or eccentric distribution families in general, we introduce a class of distribution families, *semantic information sources*, that are natural or appropriate as the object of the semantic information theory.

Although a family of distributions covers an unbounded number of distributions, the core mechanism, e.g., logic and universe, of a source to produce a family of distributions should be *bounded* due to the physical constraints. In other words, such a natural family of distributions should be actualized as an unbounded series of distributions produced by a physically bounded mechanism, e.g., logic and universe, along with an unbounded series of inputs, e.g., semantics.

Since oc-circuits are sufficiently general to express any distribution (as shown in Theorem 1), a natural and appropriate object in the semantic information theory should be expressed by an oc-circuit for a family of distributions, which is defined in Definition 4.

We have another condition for an appropriate object or its oc-circuit. Roughly, the shortest oc-circuit for simulating an appropriate object should be converging asymptotically, since we aim to characterize an object by the asymptotic properties in our theory. Then, the shortest expression (logic, universe, and semantics) of an oc-circuit simulating a family of distributions, $\mathcal{X} := \{X_n\}$, should be equivalent to the shortest one for *each* distribution X_n asymptotically (for sufficiently large n). Namely, the *global* shortest expression (the proper logic, universe, and semantics of \mathcal{X}) should be equivalent to the *local* shortest expression (the proper logic, universe, and semantics of X_n) asymptotically.

Definition 6 (*Semantic Information Source*)

A family of distributions, $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$, is called a “semantic information source” at precision level $\delta(\cdot)$ if there exists an oc-circuit for a family of distributions, $\mathcal{C}^{\mathcal{X}} := (\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, \infty, \vec{m}_{\infty}^{\mathcal{X}})$, where $\mathcal{C}_n^{\mathcal{X}} := (\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, n, \vec{m}_n^{\mathcal{X}})$, and $\vec{m}_n^{\mathcal{X}}$ is the $(N_m^{\mathcal{X}} \cdot \lceil n/L_y^{\mathcal{X}} \rceil - \text{bit})$ prefix of $\vec{m}_{\infty}^{\mathcal{X}}$ for n -bit output, that satisfies the following conditions.

- For all $n \in \mathbb{N}$, $X_n \stackrel{\delta(n)}{\approx} Y_n^{\mathcal{X}} \stackrel{\mathcal{R}}{\leftarrow} \mathcal{C}_n^{\mathcal{X}}$, and
- there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$|(\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, \vec{m}_n^{\mathcal{X}})| = \min\{ |(\overline{\mathcal{C}}, u, \vec{m}_n)| \mid X_n \stackrel{\delta(n)}{\approx} Y_n \stackrel{\mathcal{R}}{\leftarrow} (\overline{\mathcal{C}}, u, n, \vec{m}_n) \}. \quad (2)$$

If there are multiple oc-circuits, $\mathcal{C}^{\mathcal{X}}$, that satisfy the above conditions, the lexicographically first one is selected as $\mathcal{C}^{\mathcal{X}}$ for \mathcal{X} .

Then, $\mathcal{C}^{\mathcal{X}}$, $\overline{\mathcal{C}}^{\mathcal{X}}$, $u^{\mathcal{X}}$, and $\vec{m}_{\infty}^{\mathcal{X}}$ are called the “proper oc-circuit,” “proper logic,” “proper universe,” and “proper semantics” of \mathcal{X} at precision level $\delta(\cdot)$, respectively.

Here, $\mathcal{X} \stackrel{\delta(\cdot)}{\approx} \mathcal{Y} \stackrel{\mathcal{R}}{\leftarrow} \mathcal{C}^{\mathcal{X}} := (\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, \infty, \vec{m}_{\infty}^{\mathcal{X}})$.

For two semantic information sources, $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ and $\mathcal{Y} := \{Y_n\}_{n \in \mathbb{N}}$, we say \mathcal{X} and \mathcal{Y} are “semantically equivalent” at precision level $\delta(\cdot)$ iff they have the same proper oc-circuit \mathcal{C} at level $\delta(\cdot)$. We denote this by $\mathcal{X} \stackrel{\delta(\cdot)}{=} \mathcal{Y}$.

Remark 8 From the definition, for sufficiently large n ($\exists n_0 \forall n > n_0$),

$$|\mathcal{C}_n^{\mathcal{X}}| = |\mathcal{C}^{X_n}| = \text{OC}(X_n, \delta(n)),$$

where $\mathcal{C}^{X_n} := (\overline{\mathcal{C}}^{X_n}, u^{X_n}, n, \vec{m}^{X_n})$ is the proper oc-circuit of X_n at precision level $\delta(n)$ (see Definition 2 for the proper oc-circuit).

The left term in Eq. (2) is fixed by \mathcal{X} , while the right term varies with each X_n . This definition says that, in semantic information source $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$, X_n for all sufficiently large n is uniformly characterized by a single oc-circuit $\mathcal{C}^{\mathcal{X}}$ proper for \mathcal{X} .

Semantic information source $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ is characterized by its proper oc-circuit $(\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, \infty, \vec{m}_{\infty}^{\mathcal{X}})$, where $\vec{m}_{\infty}^{\mathcal{X}} := (m_i)_{i=1,2,\dots}$ with $m_i \in \{0,1\}^{N_m^{\mathcal{X}}}$. Since $\vec{m}_{\infty}^{\mathcal{X}}$ includes an infinite number of strings in $\{0,1\}^{N_m^{\mathcal{X}}}$, any value in $\{0,1\}^{N_m^{\mathcal{X}}}$ could be m_i for some $i \in \mathbb{N}$, or the value of m_i for $i \in \mathbb{N}$ could be any value in $\{0,1\}^{N_m^{\mathcal{X}}}$. Hence, if $(\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, \infty, \vec{m}_{\infty}^{\mathcal{X}})$ is the proper oc-circuit of semantic information source \mathcal{X} , an oc-circuit $(\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, \infty, \vec{m}_{\infty})$ with any other \vec{m}_{∞} could be the proper oc-circuit of a semantic information source with the proper semantics \vec{m}_{∞} .

Therefore, a semantic information source is characterized by its proper logic $\overline{\mathcal{C}} := (C, N_u, N_s, N_m, N_r, L_y, s_1)$ along with universe u in the universe space $\mathbb{U}^{\overline{\mathcal{C}}} :=$

$\{0,1\}^{N_u}$ and semantics \vec{m}_{∞} in the semantics space $\mathbb{M}^{\overline{\mathcal{C}}} := \{(m_i)_{i=1,2,\dots} \mid m_i \in \{0,1\}^{N_m}\}$.

Namely, the $L_y^{\mathcal{X}}$ bit output, y_i , should have $N_m^{\mathcal{X}}$ bit semantic information, i.e., as n bit output should have $nN_m^{\mathcal{X}}/L_y^{\mathcal{X}}$ (or its rounded-up integer, $\lceil nN_m^{\mathcal{X}}/L_y^{\mathcal{X}} \rceil$) bit semantic information.

Definition 7 (Semantic Information Amount and Semantic Information Space)

Let $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ be a semantic information source whose proper logic at precision level $\delta(\cdot)$ is $\overline{\mathcal{C}}^{\mathcal{X}} := (C^{\mathcal{X}}, N_u^{\mathcal{X}}, N_s^{\mathcal{X}}, N_m^{\mathcal{X}}, N_r^{\mathcal{X}}, L_y^{\mathcal{X}}, s_1^{\mathcal{X}})$.

The “semantic information amount,” SA, of $X_n \in \mathcal{X}$ at precision level $\delta(\cdot)$ is defined by

$$\text{SA}(X_n, \delta(\cdot)) := \lceil nN_m^{\mathcal{X}}/L_y^{\mathcal{X}} \rceil.$$

Let $\mathbb{M}^{\overline{\mathcal{C}}} := \{(m_i)_{i=1,2,\dots} \mid m_i \in \{0,1\}^{N_m}\}$ be the “semantic information (meaning) space” of proper logic $\overline{\mathcal{C}} := (C, N_u, N_s, N_m, N_r, L_y, s_1)$, and $\mathbb{M}_n^{\overline{\mathcal{C}}} := \{(m_i)_{i=1,\dots,K_n} \mid m_i \in \{0,1\}^{N_m}\}$ be “the $(N_m \cdot \lceil n/L_y^{\mathcal{X}} \rceil - \text{bit})$ prefix of $\mathbb{M}^{\overline{\mathcal{C}}}$ for an n -bit output.”

Let $\mathbb{U}^{\overline{\mathcal{C}}} := \{0,1\}^{N_u}$ be the “universe space” of proper logic $\overline{\mathcal{C}}$.

Examples of Semantic Information Sources

Here we show some examples of semantic information sources.

1. (Example 1)

Let us employ an example of Shakespeare’s plays again, and imaginarily suppose that there are an unbounded number of Shakespeare’s plays, but that the logic and universe (knowledge) of Shakespeare are bounded.

Let $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ be a sequential family of distributions of Shakespeare’s (unbounded number of) plays.

Given $X_{n^{(1)}}$ with $n^{(1)} \in \mathbb{N}$, we compute an oc-circuit $\mathcal{C}_{n^{(1)}}^{(1)}$ such that $\mathcal{C}_{n^{(1)}}^{(1)} := (\overline{\mathcal{C}}^{(1)}, u^{(1)}, n^{(1)}, \vec{m}_{n^{(1)}}^{(1)})$ is the shortest (proper) oc-circuit to simulate $X_{n^{(1)}}$ at precision level δ , i.e., $\text{OC}(X_{n^{(1)}}, \delta) = |\mathcal{C}_{n^{(1)}}^{(1)}|$.

Next, for some $n^{(2)} > n^{(1)}$, we compute the shortest (proper) oc-circuit $\mathcal{C}_{n^{(2)}}^{(2)} := (\overline{\mathcal{C}}^{(2)}, u^{(2)}, n^{(2)}, \vec{m}_{n^{(2)}}^{(2)})$ to simulate $X_{n^{(2)}}$ at precision level δ .

If for any $n^{(2)} > n^{(1)}$ $|(\overline{\mathcal{C}}^{(2)}, u^{(2)}, (\vec{m}_{n^{(2)}}^{(2)})_{n^{(1)}})| = |(\overline{\mathcal{C}}^{(1)}, u^{(1)}, \vec{m}_{n^{(1)}}^{(1)})|$ where $(\vec{m}_{n^{(2)}}^{(2)})_{n^{(1)}}$ is the $n^{(1)}$ -prefix of $\vec{m}_{n^{(2)}}^{(2)}$, it could imply that $\mathcal{C}^{(1)}$, $u^{(1)}$ and $\lim_{n^{(2)} \rightarrow \infty} \vec{m}_{n^{(2)}}^{(2)}$ are the proper logic, universe, and semantics of \mathcal{X} , respectively.

If for some $n^{(2)} > n^{(1)}$ $|(\overline{\mathcal{C}}^{(2)}, u^{(2)}, (\vec{m}_{n^{(2)}}^{(2)})_{n^{(1)}})| \neq |(\overline{\mathcal{C}}^{(1)}, u^{(1)}, \vec{m}_{n^{(1)}}^{(1)})|$, let $\mathcal{C}_{n^{(2)}}^* := \mathcal{C}_{n^{(2)}}^{(2)}$ as a candidate of the proper oc-circuit of \mathcal{X} (up to the size of $n^{(2)}$).

We repeat the procedure for $n^{(i)}$ ($i = 3, 4, \dots$) and update $\mathcal{C}_{n^{(i)}}^*$.

If $|\overline{(\mathcal{C}^{(i+1)}, u^{(i+1)}, (\vec{m}_{n^{(i+1)}}^{(i+1)})_{n^{(i)}})}| \neq |\overline{(\mathcal{C}^{(i)}, u^{(i)}, \vec{m}_{n^{(i)}}^{(i)})}|$ for some $i \in \mathbb{N}$, it should hold that $|\overline{(\mathcal{C}^{(i+1)}, u^{(i+1)})}| > |\overline{(\mathcal{C}^{(i)}, u^{(i)})}|$, since the required logic and universe (knowledge) to understand the plays should increase as the amount of plays increases.

In the updating process of $\mathcal{C}_{n^{(i)}}^*$, the i -th semantics part, $\vec{m}_{n^{(i)}}^{(i)}$, with $X_{n^{(i)}}$ has some redundancy in light of a longer (more global) context with $X_{n^{(i+1)}}$ and such redundancy could be eliminated in $(\vec{m}_{n^{(i+1)}}^{(i+1)})_{n^{(i)}}$ and absorbed into the $(i+1)$ -th logic and universe, $(\overline{(\mathcal{C}^{(i+1)}, u^{(i+1)})})$ (for a longer context), i.e., the logic and universe part should increase in the process, while the semantics part becomes more compressed (shorter) and closer to a uniform one. IN addition, block size $N_y^{(i)}$ of the output becomes longer, where a longer block with a longer context is processed using more complicated logic and a larger universe.

The logic and universe part, $(\overline{(\mathcal{C}^{(i)}, u^{(i)})})$, of oc-circuit $\mathcal{C}^{(i)}$ should be finitely bounded for any $i \in \mathbb{N}$. Actually,

$$|\overline{(\mathcal{C}^{(1)}, u^{(1)})}| \leq |\overline{(\mathcal{C}^{(2)}, u^{(2)})}| \leq \dots \leq |\overline{(\mathcal{C}^*, u^*)}|,$$

where $(\overline{(\mathcal{C}^*, u^*)})$ should be the proper logic and universe of \mathcal{X} .

Hence, there exists $i^* \in \mathbb{N}$ such that for any $i > i^*$ ($n^{(i)} > n^{(i^*)}$) $|\overline{(\mathcal{C}^{(i)}, u^{(i)}, (\vec{m}_{n^{(i)}}^{(i)})_{n^{(i^*)}})}| = |\overline{(\mathcal{C}^{(i^*)}, u^{(i^*)}, \vec{m}_{n^{(i^*)}}^{(i^*)})}|$, i.e., there exists $i^* \in \mathbb{N}$ and $(\overline{\mathcal{C}^\mathcal{X}}, u^\mathcal{X}, \vec{m}_\infty^\mathcal{X}) := (\overline{\mathcal{C}^{(i^*)}}, u^{(i^*)}, \lim_{n^{(i^*)} \rightarrow \infty} \vec{m}_{n^{(i^*)}}^{(i^*)})$ such that for any $i > i^*$ ($n^{(i)} > n^{(i^*)}$) $|\overline{(\mathcal{C}^\mathcal{X}, u^\mathcal{X}, (\vec{m}_\infty^\mathcal{X})_{n^{(i)}})}| = |\overline{(\mathcal{C}^{(i)}, u^{(i)}, (\vec{m}_{n^{(i)}}^{(i)})_{n^{(i)}})}|$.

Thus, \mathcal{X} is a semantic information source and $\mathcal{C}^\mathcal{X} := (\overline{\mathcal{C}^\mathcal{X}}, u^\mathcal{X}, \infty, \vec{m}_\infty^\mathcal{X})$ is the proper oc-circuit of \mathcal{X} .

2. (Example 2)

Let \mathcal{C} be an oc-circuit that outputs a sequential family of distributions, $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}} \stackrel{\mathcal{R}}{\leftarrow} \mathcal{C}$, such that $\mathcal{C} := (\overline{\mathcal{C}}, u, \infty, \vec{m}_\infty)$ and $\vec{m}_\infty := \{\vec{m}_n\}_{n \in \mathbb{N}}$, $\vec{m}_n \stackrel{\mathcal{U}}{\leftarrow} \{0, 1\}^{\lceil nN_m/L_y \rceil}$.

For any precision level δ , for sufficiently large $n^* \in \mathbb{N}$, we can compute $\mathcal{C}_{n^*}^* := (\overline{\mathcal{C}^*}, u^*, n^*, \vec{m}_{n^*}^*)$ which is the shortest (proper) oc-circuit to simulate X_{n^*} at precision level δ .

Then, there exists \vec{m}_∞^* with high probability such that $\mathcal{C}^* := (\overline{\mathcal{C}^*}, u^*, \infty, \vec{m}_\infty^*)$, and for any $n > n^*$ $\mathcal{C}_n^* := (\overline{\mathcal{C}^*}, u^*, \infty, (\vec{m}_\infty^*)_n)$ is the shortest oc-circuit

of X_n at precision level δ , i.e., \mathcal{C}^* is the proper oc-circuit of \mathcal{X} at precision level δ .

This is because $\vec{m}_n \stackrel{\mathcal{U}}{\leftarrow} \{0, 1\}^{\lceil nN_m/L_y \rceil}$ and no more data compression on \vec{m}_n is possible for any sufficiently large $n > n^*$ with high probability.

That is, \mathcal{X} is a semantic information source with high probability.

The difference between this example and the first example is that the unbounded semantics sequence, \vec{m}_∞ , in this example is uniformly selected from the beginning, while, in the first example, the semantics sequence is gradually compressed as size n of distribution X_n becomes longer in the process of updating \mathcal{C}_n^* .

3. (Other Examples)

The information sources modeled in the previous semantic information theories in literature (in Section III A) are considered to be “semantic information sources.”

For example in [2], Fig. 2 shows a model of semantic information communication. Here, I_S (Inference Procedure) and the syntax and logic part of M_S (Message generator) can be considered as *circuit* \mathcal{C} of the oc-circuit and W_S (world model), K_S (Background Knowledge) can be considered as *universe* u of the oc-circuit, and the semantics of M_S (Message generator) can be considered as semantics \vec{m} of the oc-circuit. That is, the messages from Sender S in Fig. 2 can be modeled as a source generated by an oc-circuit, or a semantic information source.

We then consider the following problem. Given semantic information source \mathcal{X} with $\delta(\cdot)$, can we compute its proper oc-circuit and the related information? The answer is no, since \mathcal{X} consists of an infinite number of distributions and it cannot be described finitely.

The following theorem however, shows that, given $X_n \in \mathcal{X}$ and $\delta(\cdot)$ with a sufficiently large n , we can compute the proper oc-circuit of \mathcal{X} .

Theorem 4 *For any semantic information source \mathcal{X} at precision level $\delta(\cdot) > 0$, given $X_n \in \mathcal{X}$ and $\delta(n)$ for some $n > n_0$, where n_0 is given in Definition 6, the proper oc-circuit (proper logic, proper universe, and n -prefix of proper semantics) of \mathcal{X} at precision level $\delta(\cdot)$ can be computed.*

The proof of this theorem is essentially the same as that for Theorem 1.

Remark 9 *In our semantic information theory, the concepts of “proper oc-circuit,” “proper logic,” “proper universe,” and “proper semantics” introduced in Definition 6 and the computability shown in Theorem 4 represent a positive answer to the questions raised at the end of Section III A.*

We then introduce the concept of *conditional oc-circuit*, *conditional semantic information source* and *conditional semantic information amount*, which play central roles in our theory, especially in the semantic channel coding theorem (Section III B 3), the effectiveness coding theorem (Section III B 4) and in the semantic source coding theorem (Section III B 2).

Definition 8 (Conditional OC-Circuits)

Let $\mathcal{Z} := \{Z_n\}_{n \in \mathbb{N}}$ be a sequential family of distributions. A “conditional oc-circuit for a family of distributions under \mathcal{Z} ” is $\mathcal{C}^{\mathcal{Z}} := (\bar{\mathcal{C}}^{\mathcal{Z}}, \infty, u^{\mathcal{Z}}, \vec{m}_\infty^{\mathcal{Z}})$, where

$$\bar{\mathcal{C}}^{\mathcal{Z}} := (C^{\mathcal{Z}}, N_u^{\mathcal{Z}}, N_s^{\mathcal{Z}}, N_m^{\mathcal{Z}}, N_r^{\mathcal{Z}}, N_z^{\mathcal{Z}}, L_y^{\mathcal{Z}}, s_1^{\mathcal{Z}}),$$

$$\vec{m}_\infty^{\mathcal{Z}} := (m_i)_{i=1,2,\dots} := (m_1, m_2, \dots) \in \{0,1\}^\infty,$$

$$(z_1, z_2, \dots) \stackrel{R}{\leftarrow} \mathcal{Z}$$

$$(s_{i+1}, y_i) \leftarrow \boxed{C^{\mathcal{Z}}(u^{\mathcal{Z}}, \cdot)} \leftarrow \begin{array}{c} \mathcal{Z} \\ \downarrow \\ \boxed{z_i}, s_i, m_i, r_i \end{array}, \quad i = 1, 2, \dots,$$

$$\text{i.e., } (s_{i+1}, y_i) := C^{\mathcal{Z}}(u^{\mathcal{Z}}, z_i, s_i, m_i, r_i), \quad i = 1, 2, \dots,$$

$$Y_n := (y_1, \dots, y_{K_n^{\mathcal{Z}}})_n, \quad K_n^{\mathcal{Z}} := \lceil n/L_y^{\mathcal{Z}} \rceil$$

$$\mathcal{Y} := \{Y_n\}_{n \in \mathbb{N}} \stackrel{R}{\leftarrow} \mathcal{C}^{\mathcal{Z}},$$

where $m_i \in \{0,1\}^{N_m^{\mathcal{Z}}}$, $z_i \in \{0,1\}^{N_z^{\mathcal{Z}}}$, $s_i \in \{0,1\}^{N_s^{\mathcal{Z}}}$, $u^{\mathcal{Z}} \in \{0,1\}^{N_u^{\mathcal{Z}}}$, $r_i \stackrel{U}{\leftarrow} \{0,1\}^{N_r^{\mathcal{Z}}}$, and $y_i \in \{0,1\}^{L_y^{\mathcal{Z}}}$. The probability on \mathcal{Y} is taken over the randomness of \mathcal{Z} and $\{r_i\}$,

Remark 10 Given a sample of sequential family of distributions \mathcal{Z} , conditional oc-circuit $\mathcal{C}^{\mathcal{Z}}$ divides the sample of \mathcal{Z} into $N_z^{\mathcal{Z}}$ -bit strings, z_1, z_2, \dots , where the size, $N_z^{\mathcal{Z}}$, is also determined by $\mathcal{C}^{\mathcal{Z}}$.

Definition 9 (Conditional Semantic Information Source and Conditional Semantic Information Amount)

Let \mathcal{Z} be a sequential family of distributions. A family of distributions, $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$, is called a “conditional semantic information source under \mathcal{Z} ” at precision level $\delta(\cdot)$ if there exists an conditional oc-circuit for a family of distributions under \mathcal{Z} , $\mathcal{C}^{\mathcal{X}:\mathcal{Z}} := (\bar{\mathcal{C}}^{\mathcal{X}:\mathcal{Z}}, u^{\mathcal{X}:\mathcal{Z}}, \infty, \vec{m}_\infty^{\mathcal{X}:\mathcal{Z}})$, where $\mathcal{C}_n^{\mathcal{X}:\mathcal{Z}} := (\bar{\mathcal{C}}^{\mathcal{X}:\mathcal{Z}}, u^{\mathcal{X}:\mathcal{Z}}, n, \vec{m}_n^{\mathcal{X}:\mathcal{Z}})$ and $\vec{m}_n^{\mathcal{X}:\mathcal{Z}}$ is the $(N_m^{\mathcal{X}:\mathcal{Z}} \cdot \lceil n/L_y^{\mathcal{X}:\mathcal{Z}} \rceil)$ -bit prefix of $\vec{m}_\infty^{\mathcal{X}:\mathcal{Z}}$ for n -bit output, that satisfies the following conditions.

- For all $n \in \mathbb{N}$, $X_n \stackrel{\delta(n)}{\approx} Y_n^{\mathcal{X}:\mathcal{Z}} \stackrel{R}{\leftarrow} \mathcal{C}_n^{\mathcal{X}:\mathcal{Z}}$ with any sampled value of $Z_{N_z^{\mathcal{X}:\mathcal{Z}} K_n^{\mathcal{X}:\mathcal{Z}}} \stackrel{R}{\leftarrow} \mathcal{Z}$, and
- there exists n_0 such that for all $n > n_0$,

$$|(\bar{\mathcal{C}}^{\mathcal{X}:\mathcal{Z}}, u^{\mathcal{X}:\mathcal{Z}}, \vec{m}_n^{\mathcal{X}:\mathcal{Z}})| = \min\{ |(\bar{\mathcal{C}}^{\mathcal{Z}}, u^{\mathcal{Z}}, \vec{m}_n^{\mathcal{Z}})| \mid$$

$$X_n \stackrel{\delta(n)}{\approx} Y_n^{\mathcal{Z}} \stackrel{R}{\leftarrow} (\bar{\mathcal{C}}^{\mathcal{Z}}, u^{\mathcal{Z}}, n, \vec{m}_n^{\mathcal{Z}})$$

with any sampled value of $Z_{N_z^{\mathcal{Z}} K_n^{\mathcal{Z}}} \stackrel{R}{\leftarrow} \mathcal{Z}$. (3)

If there are multiple conditional oc-circuits, $\mathcal{C}^{\mathcal{X}:\mathcal{Z}}$, that satisfy the above conditions, the lexicographically first one is selected as $\mathcal{C}^{\mathcal{X}:\mathcal{Z}}$.

Then, the “conditional semantic information amount,” SA, of $X_n \in \mathcal{X}$ under \mathcal{Z} at precision level $\delta(n)$ is

$$\text{SA}(X_n : \mathcal{Z}, \delta(n)) := \lceil n N_m^{\mathcal{X}:\mathcal{Z}} / L_y^{\mathcal{X}:\mathcal{Z}} \rceil,$$

where $\bar{\mathcal{C}}^{\mathcal{X}:\mathcal{Z}} := (C^{\mathcal{X}:\mathcal{Z}}, N_u^{\mathcal{X}:\mathcal{Z}}, N_s^{\mathcal{X}:\mathcal{Z}}, N_m^{\mathcal{X}:\mathcal{Z}}, N_r^{\mathcal{X}:\mathcal{Z}}, N_z^{\mathcal{X}:\mathcal{Z}}, L_y^{\mathcal{X}:\mathcal{Z}}, s_1^{\mathcal{X}:\mathcal{Z}})$.

Remark 11 Eq.(3) can be written as below in a manner similar to that for Eq.(2) shown in Remark 8. For sufficiently large n ,

$$|\mathcal{C}_n^{\mathcal{X}:\mathcal{Z}}| = |\mathcal{C}^{X_n:\mathcal{Z}}|,$$

where $\mathcal{C}^{X_n:\mathcal{Z}} := (\bar{\mathcal{C}}^{X_n:\mathcal{Z}}, u^{X_n:\mathcal{Z}}, n, \vec{m}_n^{X_n:\mathcal{Z}})$ is the shortest one in $\{(\bar{\mathcal{C}}^{\mathcal{Z}}, u^{\mathcal{Z}}, n, \vec{m}_n^{\mathcal{Z}}) \mid X_n \stackrel{\delta(n)}{\approx} Y_n^{\mathcal{Z}} \stackrel{R}{\leftarrow} (\bar{\mathcal{C}}^{\mathcal{Z}}, u^{\mathcal{Z}}, n, \vec{m}_n^{\mathcal{Z}}) \text{ with any sampled value of } Z_{N_z^{\mathcal{Z}} K_n^{\mathcal{Z}}} \stackrel{R}{\leftarrow} \mathcal{Z}\}$.

Based on the conditional semantic information amount, we next introduce the concept of the *semantic mutual information amount*, which is employed in the effectiveness to be shown in Section III B 4.

Definition 10 (Semantic Mutual Information Amount)

Let $\mathcal{Z} := \{Z_n\}_{n \in \mathbb{N}}$ be a sequential family of distributions and $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ be a semantic information source and a conditional semantic information source under \mathcal{Z} .

The “semantic mutual information amount” of distribution $X_n \in \mathcal{X}$ with \mathcal{Z} , $\text{SI}(X_n : \mathcal{Z}, \delta(n))$, is defined by

$$\text{SI}(X_n : \mathcal{Z}, \delta(n)) := \text{SA}(X_n, \delta(n)) - \text{SA}(X_n, \mathcal{Z}, \delta(n)).$$

Remark 12 Semantic mutual information amount $\text{SI}(X_n : \mathcal{Z}, \delta(n))$ means the semantic information amount in \mathcal{Z} with respect to X_n . More precisely, it means the semantic information amount in $Z_{\ell(n)}$ with respect to X_n , where $\ell(n) := N_z^{\mathcal{X}:\mathcal{Z}} \cdot \lceil n/L_y^{\mathcal{X}:\mathcal{Z}} \rceil$.

In contrast to the mutual information amount in the Shannon information theory, the semantic mutual information amount is not symmetric, i.e., $\text{SI}(X_n : \mathcal{Z}, \delta(n))$ is not always equivalent to $\text{SI}(Z_{\ell(n)} : \mathcal{X}, \delta(n))$ for some $\ell(\cdot)$, since even if some semantic information X is useful for Z , Z may not be so useful for X , e.g., quantum physics is often useful to understand chemical phenomena but the converse is not always true.

2. Semantic Source Coding

Based on the notion of (conditional) semantic information sources, we next develop a semantic information theory for Level B (semantic) problem as described by Weaver [41].

Our theory answers two fundamental questions in semantic information theory: What is the ultimate semantic data compression, or ultimate data compression with preserving semantics, and what is the ultimate transmission rate of semantic data communication. The first question is answered by Theorem 5, *semantic source coding theorem*, in this section, and the second question is answered by Theorems 6 and 7, *semantic channel coding theorem*, in Section III B 3.

To begin with, we introduce the notion of semantic data compression in the following definition.

Definition 11 (*Semantic Source Coding System*)

Let $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ be a semantic information source. Semantic source coding system $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ consists of sender \mathcal{S} , which outputs a sequential family of distributions, \mathcal{Z} , on \mathcal{X} , and receiver \mathcal{R} , which is a conditional oc-circuit $\mathcal{C}^{\mathcal{Z}}$ under \mathcal{Z} without semantics input ($N_m^{\mathcal{Z}} = 0$) to output a family of distributions \mathcal{Y} , where $\mathcal{C}^{\mathcal{Z}} := (\overline{\mathcal{C}}^{\mathcal{Z}}, u^{\mathcal{Z}}, \infty, \vec{m}_\infty^{\mathcal{Z}})$, and $\overline{\mathcal{C}}^{\mathcal{Z}} := (C^{\mathcal{Z}}, N_u^{\mathcal{Z}}, N_s^{\mathcal{Z}}, N_m^{\mathcal{Z}}, N_r^{\mathcal{Z}}, N_z^{\mathcal{Z}}, L_y^{\mathcal{Z}}, s_1^{\mathcal{Z}})$.

For parameter $n \in \mathbb{N}$, sender \mathcal{S} outputs $Z_{\ell(n)}$ on $X_n \in \mathcal{X}$, which is directly input to receiver \mathcal{R} , and \mathcal{R} outputs $Y_n \stackrel{\mathcal{R}}{\leftarrow} \mathcal{R}(n, Z_{\ell(n)})$, where $\ell(n) := N_z^{\mathcal{Z}} \cdot \lceil n/L_y^{\mathcal{Z}} \rceil$, $Z_{\ell(n)} \in \mathcal{Z}$ is a distribution over $\{0, 1\}^{\ell(n)}$, and $Y_n \in \mathcal{Y}$ is a distribution over $\{0, 1\}^n$.

$$\boxed{\mathcal{S}^{\mathcal{X}}(n)} \xrightarrow{Z_{\ell(n)}} \boxed{\mathcal{R}(n, \cdot)} \rightarrow Y_n$$

We call $\ell(n)$ the code length of $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ in $n \in \mathbb{N}$.

We say that $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ correctly codes at precision level $\delta(\cdot)$ if there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $X_n \stackrel{\delta(n)}{\approx} Y_n$ with any sampled value of $Z_{\ell(n)} \stackrel{\mathcal{R}}{\leftarrow} \mathcal{Z}$.

The following theorem answers the above-mentioned first question. Roughly, the ultimate compression size of semantic information source $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ at precision level $\delta(\cdot)$ is its semantic information amount, $\text{SA}(X_n, \delta(n))$, asymptotically. The compression size of $\text{SA}(X_n, \delta(n)) + \epsilon$ (ϵ is a positive small value) is possible, but the compression shorter than $\text{SA}(X_n, \delta(n)) - \epsilon$ is impossible.

Theorem 5 (*Semantic Source Coding Theorem*)

Let $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ be a semantic information source with precision level $\delta(\cdot)$.

There exists a semantic source coding system $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ with code length $\ell(\cdot)$ that correctly codes at precision level $\delta(\cdot)$ and $\ell(\cdot)$ satisfies the following inequality. For any $\epsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$\text{SA}(X_n, \delta(n)) \leq \ell(n) < \text{SA}(X_n, \delta(n)) + \epsilon.$$

There is no semantic source coding system $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ such that $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ with code length $\ell(\cdot)$ correctly codes at precision level $\delta(\cdot)$ and $\ell(\cdot)$ satisfies the following inequality. For any $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ for all $n > n_0$,

$$\ell(n) < \text{SA}(X_n, \delta(n)) - \epsilon.$$

Proof

Since $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ is a semantic information source, the proper oc-circuit of \mathcal{X} , $\mathcal{C}^{\mathcal{X}} := (\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, \infty, \vec{m}_\infty^{\mathcal{X}})$, exists, where $\mathcal{C}^{\mathcal{X}} := (\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, n, \vec{m}_n^{\mathcal{X}})$ and $\overline{\mathcal{C}}^{\mathcal{X}} := (C^{\mathcal{X}}, N_u^{\mathcal{X}}, N_s^{\mathcal{X}}, N_m^{\mathcal{X}}, N_r^{\mathcal{X}}, L_y^{\mathcal{X}}, s_1^{\mathcal{X}})$.

Then, $\text{SA}(X_n, \delta(n)) := \lceil n N_m^{\mathcal{X}} / L_y^{\mathcal{X}} \rceil$ ($\approx \lceil n / L_y^{\mathcal{X}} \rceil N_m^{\mathcal{X}} = |\vec{m}_n^{\mathcal{X}}|$), where $\text{SA}(X_n, \delta(n)) \approx |\vec{m}_n^{\mathcal{X}}|$ means $\forall \epsilon > 0 \exists n_0 \forall n > n_0 \text{SA}(X_n, \delta(n)) \leq |\vec{m}_n^{\mathcal{X}}| < \text{SA}(X_n, \delta(n)) + \epsilon$.

We construct semantic source coding system $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ such that sender \mathcal{S} sends $\vec{m}_n^{\mathcal{X}}$ (as $Z_{\ell(n)}$) to receiver \mathcal{R} , i.e., $\ell(n) := |\vec{m}_n^{\mathcal{X}}| \approx \text{SA}(X_n, \delta(n))$, and \mathcal{R} is a conditional oc-circuit under \mathcal{Z} without semantics input, $\mathcal{C}^{\mathcal{Z}} := (\overline{\mathcal{C}}^{\mathcal{Z}}, u^{\mathcal{Z}}, \infty, \lambda)$, where $u^{\mathcal{Z}} := u^{\mathcal{X}}$, and $\overline{\mathcal{C}}^{\mathcal{Z}} := (C^{\mathcal{Z}}, N_u^{\mathcal{Z}}, N_s^{\mathcal{Z}}, 0, N_r^{\mathcal{Z}}, N_m^{\mathcal{Z}}, L_y^{\mathcal{Z}}, s_1^{\mathcal{Z}})$ as $(C^{\mathcal{Z}}, N_u^{\mathcal{Z}}, N_s^{\mathcal{Z}}, N_m^{\mathcal{Z}}, N_r^{\mathcal{Z}}, N_z^{\mathcal{Z}}, L_y^{\mathcal{Z}}, s_1^{\mathcal{Z}})$. Here, $\overline{\mathcal{C}}^{\mathcal{Z}}$ is the same functionality as that of $\overline{\mathcal{C}}^{\mathcal{X}}$ except the input place such that $\vec{m}_n^{\mathcal{X}}$ sent from \mathcal{S} is input to $\overline{\mathcal{C}}^{\mathcal{Z}}$ as $Z_{\ell(n)}$, i.e., $N_z^{\mathcal{Z}} := N_m^{\mathcal{X}}$, and no semantics is input to $\overline{\mathcal{C}}^{\mathcal{Z}}$, i.e., $N_m^{\mathcal{Z}} := 0$, while $\vec{m}_n^{\mathcal{X}}$ is input to $\overline{\mathcal{C}}^{\mathcal{X}}$ as the semantics.

From the definition of the proper oc-circuit, for sufficiently large n ($\exists n_0 \forall n > n_0$),

$$X_n \stackrel{\delta(n)}{\approx} Y_n \stackrel{\mathcal{R}}{\leftarrow} (\overline{\mathcal{C}}^{\mathcal{X}}, u^{\mathcal{X}}, n, \vec{m}_n^{\mathcal{X}}) = (\overline{\mathcal{C}}^{\mathcal{Z}}, u^{\mathcal{X}}, n, \lambda).$$

That is, for sufficiently large n , $X_n \stackrel{\delta(n)}{\approx} Y_n \stackrel{\mathcal{R}}{\leftarrow} \mathcal{R}(n, Z_{\ell(n)})$, i.e., the constructed $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ correctly codes at precision level $\delta(\cdot)$.

Since $\ell(n) \approx \text{SA}(X_n, \delta(n))$, $\forall \epsilon > 0 \exists n_0 \forall n > n_0$

$$\text{SA}(X_n, \delta(n)) \leq \ell(n) < \text{SA}(X_n, \delta(n)) + \epsilon.$$

This completes the former statement of this theorem.

To prove the latter statement of this theorem by contradiction, let us assume that there exists semantic source coding system $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ such that $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ correctly codes at precision level $\delta(\cdot)$ and its code length $\ell(n)$ is that $\forall \epsilon > 0, \exists n_0 \in \mathbb{N}, \forall n > n_0, \ell(n) < \text{SA}(X_n, \delta(n)) - \epsilon$.

Since \mathcal{Z} is a sequential family of distributions and $\text{SSoC}(\mathcal{S}, \mathcal{R}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ correctly codes at precision level $\delta(\cdot)$ with any sampled value of \mathcal{Z} , there exists $z_{\ell(n)} \in \{0, 1\}^{\ell(n)}$ such that $z_{\ell(n)}$ is the $\ell(n)$ -bit prefix of a sampled value of \mathcal{Z} , and for sufficiently large n , $X_n \stackrel{\delta(n)}{\approx} Y_n \stackrel{\mathcal{R}}{\leftarrow} \mathcal{R}(n, z_{\ell(n)})$.

As shown in the proof of the former statement, $\mathcal{R}(n, z_{\ell(n)})$, i.e., conditional oc-circuit $\overline{\mathcal{C}}^{\mathcal{Z}}$ with a sampled value of $\{z_{\ell(n)}\}$ and no semantics input is the same

functionality as oc-circuit \overline{C} with semantics input $\{z_{\ell(n)}\}$. That is, there exists an oc-circuit $(\overline{C}, u, n, z_{\ell(n)})$ such that for sufficiently large n ,

$$X_n \stackrel{\delta(n)}{\approx} Y_n \stackrel{R}{\leftarrow} (\overline{C}, u, n, z_{\ell(n)}).$$

Since for any ϵ and sufficiently large n $\epsilon < \text{SA}(X_n, \delta) - \ell(n)$ and $|\overline{C}, u|$ is bounded, it holds that for sufficiently large n

$$|\overline{C}, u| - |\overline{C}^{\mathcal{X}}, u^{\mathcal{X}}| < \text{SA}(X_n, \delta) - \ell(n) \leq |\vec{m}_n^{\mathcal{X}}| - |z_{\ell(n)}|,$$

where $(\overline{C}^{\mathcal{X}}, u^{\mathcal{X}}, \vec{m}_n^{\mathcal{X}})$ is the proper oc-circuit of $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$. That is,

$$|\overline{C}, u, z_{\ell(n)}| < |(\overline{C}^{\mathcal{X}}, u^{\mathcal{X}}, \vec{m}_n^{\mathcal{X}})|.$$

It contradicts the minimality of $|\overline{C}^{\mathcal{X}}, u^{\mathcal{X}}, \vec{m}_n^{\mathcal{X}}|$ and completes the proof of the latter statement. \square

Remark 13 If semantic information source $\mathcal{X} := \{X_n\}_{n \in \mathbb{N}}$ is a family of uniform distributions, its semantic information amount is zero, or $\text{SA}(X_n, \delta(n)) = 0$ for any $n \in \mathbb{N}$ and $\delta(\cdot)$. Hence, semantically compressed data size, $\ell(n)$, of X_n can be almost 0, due to Theorem 5.

It is highly contrast to the data compression capability in the traditional (Shannon) information theory: the above-mentioned source, \mathcal{X} , cannot be compressed any more because its Shannon entropy is the maximum.

As shown in this example, the semantic data compression should be more capable than the traditional data compression in many applications. That is, the semantic data compression indicated by Theorem 5 offers a great potential in various practical applications of data compression.

3. Semantic Channel Coding

In this section, we answer the second question described in the beginning of Section III B 2: What is the ultimate transmission rate of semantic data communication.

First, in the following definition, we introduce a model of semantic communication channel. Here, the semantic communication channel may be noisy or the received data from the channel may contain semantic errors. Various types of semantics errors or noises are investigated in [2]. The notion of a semantic channel coding system is introduced to correct such errors in the semantic information space over the communication channel.

Definition 12 (Semantic Channel Coding System)

Semantic channel coding system $\text{SchC} := \text{SchC}(\overline{S}^{\overline{C}, u, \text{code}}, R, \text{Ch}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ consists of sender S , which has a coding machine, $(\overline{C}, u, \text{code})$, and outputs

a sequential family of distributions, \mathcal{X} ; communication channel Ch , which receives \mathcal{X} and outputs a sequential family of distributions, \mathcal{Z} ; and receiver R , which is a conditional oc-circuit under \mathcal{Z} and outputs \mathcal{Y} .

Here, \mathcal{X} is generated by an oc-circuit, $\overline{C} := (\overline{C}, u, \infty, \vec{m}_\infty)$, where $\overline{C} := (C, N_u, N_s, N_m, N_r, L_y, s_1)$ (logic), $u \in \{0, 1\}^{N_u}$ (universe space), $\vec{m}_\infty \in \mathbb{M}^{\overline{C}} := \{(m_i)_{i=1,2,\dots} \mid m_i \in \{0, 1\}^{N_m}\}$ (semantic information space), and $\mathbb{M}^{\overline{C}} = \{0, 1\}^{N_m K_n}$ (the $N_m K_n$ -bit prefix of $\mathbb{M}^{\overline{C}}$ for n -bit output) ($K_n := \lceil n/L_y \rceil$). A coding, code, with n is $\text{code}_n : \{0, 1\}^{k(n)} \rightarrow \mathbb{M}_n^{\overline{C}}$.

For parameter $n \in \mathbb{N}$, given $\vec{m}_n^+ \in \{0, 1\}^{k(n)}$, S computes $\vec{m}_n := \text{code}_n(\vec{m}_n^+) \in \mathbb{M}_n^{\overline{C}}$ and $X_n \stackrel{R}{\leftarrow} (\overline{C}, u, n, \vec{m}_n)$, where $X_n \in \mathcal{X}$ is a distribution over $\{0, 1\}^n$. X_n is input to channel Ch , and Ch outputs $Z_{\ell(n)} \in \mathcal{Z}$, where $Z_{\ell(n)}$ is a distribution over $\{0, 1\}^{\ell(n)}$. Receiver R (conditional oc-circuit under \mathcal{Z}) receives $Z_{\ell(n)}$ and outputs $Y_n \in \mathcal{Y}$, where Y_n is a distribution over $\{0, 1\}^n$.

$$\boxed{S^{\overline{C}, u, \text{code}}(n, \vec{m}_n^+)} \xrightarrow{X_n} \boxed{\text{Ch}(n, \cdot)} \xrightarrow{Z_{\ell(n)}} \boxed{R(n, \cdot)} \rightarrow Y_n$$

We say that $\text{SchC} := \text{SchC}(\overline{S}^{\overline{C}, u, \text{code}}, R, \text{Ch}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ correctly codes at precision level $\delta(\cdot)$ if there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, for all $\vec{m}_n^+ \in \{0, 1\}^{k(n)}$, $X_n \stackrel{\delta(n)}{\approx} Y_n$ with any sampled value of $Z_{\ell(n)} \stackrel{R}{\leftarrow} \mathcal{Z}$.

We next define the semantic channel capacity of a semantic channel and semantic communication rate of a semantic coding. Roughly, the semantic channel capacity represents the maximum rate of semantic information that can be transmitted over the semantic channel, or the ratio of the maximum semantic communication amount (size) to the communication data size. The semantic communication rate is the ratio of semantic information size (input size to the coding) to the communication data size, $k(n)/n$, in the semantic coding.

Definition 13 (Semantic Channel Capacity and Semantic Communication Rate)

Let $\text{Sch} := (S^{\overline{C}, u}, \text{Ch}; \mathcal{X}, \mathcal{Z})$ be a “semantic channel” of semantic channel coding system $\text{SchC}(\overline{S}^{\overline{C}, u, \text{code}}, R, \text{Ch}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$. Given $\vec{m}_\infty \in \mathbb{M}^{\overline{C}}$, S in Sch computes $\mathcal{X} := \{X(\vec{m}_n) := X_n \stackrel{R}{\leftarrow} (\overline{C}, u, n, \vec{m}_n)\}_{n \in \mathbb{N}}$, i.e., $\mathcal{X} \stackrel{R}{\leftarrow} (\overline{C}, u, \infty, \vec{m}_\infty)$, where \vec{m}_n is the $N_m \cdot \lceil n/L_y \rceil$ -bit prefix of \vec{m}_∞ for n -bit output. Then, $\mathcal{X} := \{X(\vec{m}_n)\}_{n \in \mathbb{N}}$ is input to channel Ch in Sch , and Ch outputs $\mathcal{Z} := \{Z_{\ell(n)}\}_{n \in \mathbb{N}}$.

If for any $\vec{m}_\infty \in \mathbb{M}^{\overline{C}}$, $\mathcal{X} \stackrel{R}{\leftarrow} (\overline{C}, u, \infty, \vec{m}_\infty)$ is a conditional semantic information source under \mathcal{Z} in semantic channel Sch , we call Sch “normal”.

Let $\mathcal{M} := \{M_n\}_{n \in \mathbb{N}}$ be a sequential family of distributions over $\mathbb{M}^{\overline{C}}$, where M_n is a distribution over $\mathbb{M}_n^{\overline{C}}$, $n \in \mathbb{N}$, i.e., $M_n := \{(\vec{m}_n, p_{\vec{m}_n}) \mid \vec{m}_n \in \mathbb{M}_n^{\overline{C}}\}$, (see Definition 5 for the sequential family of distributions).

When semantic channel SCh is normal, “semantic channel capacity” SC of SCh for $n \in \mathbb{N}$ (say SCh_n) is \square

$$\text{SC}(\text{SCh}_n, \delta(n)) :=$$

$$\frac{1}{n} \cdot \max_{M_n \in \mathcal{M}_n^{\text{seq}}} \{H(M_n) - \mathbb{E}_{M_n}(\text{SA}(X(\vec{m}_n) : \mathcal{Z}, \delta(n))\}, \quad (4)$$

where $\mathcal{M}_n^{\text{seq}}$ is the class of the $N_m \cdot \lceil n/L_y \rceil$ -bit prefix of sequential families of distributions, i.e., $M_n \in \mathcal{M}_n^{\text{seq}}$ is the $N_m \cdot \lceil n/L_y \rceil$ -bit prefix of a sequential family of distributions, $H(\cdot)$ is the Shannon entropy and $\mathbb{E}_{M_n}(\cdot)$ is the expectation value over the distribution of $\vec{m}_n \stackrel{\text{R}}{\leftarrow} M_n \in \mathcal{M}_n^{\text{seq}}$.

“Semantic communication rate” SR of semantic coding $(\vec{C}, u, \text{code})$ in semantic channel coding system $\text{SChC}(\text{S}(\vec{C}, u, \text{code}), R, \text{Ch}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ for $n \in \mathbb{N}$ is

$$\text{SR}(\text{code}_n) := k(n)/n. \quad (5)$$

We say a semantic channel coding system, SChC, is “normal,” if the semantic channel, SCh, of SChC is normal.

We now show that the semantic capacity is the upper limit (or theoretically maximum) rate of semantic data transmission at which we can send semantic information over the semantic channel and recover the information at the output in the semantic channel coding system.

Theorem 6 (Semantic Channel Coding Theorem (1))

There exists no normal semantic channel coding system $\text{SChC}(\text{S}(\vec{C}, u, \text{code}), R, \text{Ch}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ that correctly codes at precision level $\delta(\cdot)$ and there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$\text{SC}(\text{SCh}_n, \delta(n)) < \text{SR}(\text{code}_n).$$

Proof

To prove this theorem by contradiction, we first assume that SChC correctly codes at precision level $\delta(\cdot)$ while satisfying $\text{SR}(\text{code}_n) > \text{SC}(\text{SCh}_n, \delta(n))$.

We then construct a distribution, $M_n^+ := \{(\vec{m}_n, p_{\vec{m}_n})\}$, such that $p_{\vec{m}_n} := 1/2^{k(n)}$ if $\vec{m}_n := \text{code}_n(\vec{m}_n^+)$ with $\vec{m}_n^+ \in \{0, 1\}^{k(n)}$, and $p_{\vec{m}_n} := 0$ otherwise. Clearly, $H(M_n^+) = k(n) = n \cdot \text{SR}(\text{code}_n)$.

Since SChC correctly codes at precision level $\delta(\cdot)$ for any $\vec{m}_n := \text{code}_n(\vec{m}_n^+)$ which occurs with probability 1 in M_n^+ , R of SChC, a conditional oc-circuit under \mathcal{Z} , outputs $Y_n \stackrel{\delta(n)}{\approx} X(\vec{m}_n)$. That is, $\text{SA}(X(\vec{m}_n) : \mathcal{Z}, \delta(n)) = 0$ for any \vec{m}_n that occurs in M_n^+ with non-zero probability. Therefore, $\mathbb{E}_{M_n^+}(\text{SA}(X(\vec{m}_n) : \mathcal{Z}, \delta(n))) = 0$.

Due to the definition (maximality) of $\text{SC}(\text{SCh}_n, \delta(n))$, we obtain that $\text{SC}(\text{SCh}_n, \delta(n)) \geq \frac{1}{n} \cdot (H(M_n^+) - \mathbb{E}_{M_n^+}(\text{SA}(X(\vec{m}_n) : \mathcal{Z}, \delta(n))) = \frac{1}{n} \cdot (n \cdot \text{SR}(\text{code}_n)) = \text{SR}(\text{code}_n)$.

This contradicts the assumption that $\text{SR}(\text{code}_n) > \text{SC}(\text{SCh}_n, \delta(n))$.

In the theorem above, it is shown that a semantic data transmission rate over a channel is impossible beyond the semantic capacity of the channel. We next present that a semantic transmission rate slightly below the semantic capacity is possible over a class of semantic channels, uniform semantic channels.

Definition 14 (Uniform Semantic Channel)

Let M_n^* be the value (distribution) of $M_n \in \mathcal{M}_n^{\text{seq}}$ to maximize Eq.(4) and $\{(\vec{m}_n, p_{\vec{m}_n}^*) \mid \vec{m}_n \in \mathbb{M}_n^{\vec{C}}\} := M_n^*$.

We say a normal semantic channel, SCh, is “uniform” if the following conditions hold.

- (Consistency) $\mathcal{M}^* := \{M_n^*\}_{n \in \mathbb{N}}$ is a sequential family of distributions.
- (Uniformity of \mathcal{M}^*)

For any $\epsilon_1, \epsilon_2 > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$

$$\Pr[|-\log_2 p_{\vec{m}_n}^* - H(M_n^*)| > \epsilon_1] < \epsilon_2,$$

where the probability is taken over the randomness of $\vec{m}_n \stackrel{\text{R}}{\leftarrow} M_n^*$.

- (Uniformity of conditional SA)

For any $\epsilon_1, \epsilon_2 > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$

$$\Pr[|\text{SA}(X(\vec{m}_n) : \mathcal{Z}, \delta(n)) - T^*| > \epsilon_1] < \epsilon_2,$$

where the probability is taken over the randomness of $\vec{m}_n \stackrel{\text{R}}{\leftarrow} M_n^*$, and $T^* := \mathbb{E}_{M_n^*}(\text{SA}(X(\vec{m}_n) : \mathcal{Z}, \delta(n)))$.

We say a semantic channel coding system, SChC, is “uniform,” if the semantic channel, SCh, of SChC is uniform.

Theorem 7 (Semantic Channel Coding Theorem (2))

There exists a uniform semantic channel coding system, $\text{SChC}(\text{S}(\vec{C}, u, \text{code}), R, \text{Ch}; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$, that correctly codes at precision level $\delta(\cdot)$, and for any ϵ ($0 < \epsilon$), there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$\text{SC}(\text{SCh}_n, \delta(n)) - \epsilon < \text{SR}(\text{code}_n) < \text{SC}(\text{SCh}_n, \delta(n)).$$

Proof

Let M_n^* be the maximum value of M_n to maximize Eq.(4) of normal channel Ch on $(\vec{C}, u, \mathbb{M}^{\vec{C}})$ as described in Definition 13.

For $\epsilon_1 > 0$ and M_n^* , let $\mathcal{M}_{\epsilon_1, n}^* := \{\mu \mid (\mu, p_\mu^*) \in M_n^* \wedge |-\log_2 p_\mu^* - H(M_n^*)| \leq \epsilon_1 \wedge |\text{SA}(X(\mu) : \mathcal{Z}, \delta(n)) - T^*| \leq \epsilon_1\}$.

From the uniformity of \mathcal{M}^* and uniformity of the conditional SA, it holds that $\forall \epsilon_1, \epsilon_2 > 0, \exists n_0 \in \mathbb{N}, \forall n > n_0$,

$\Pr[\mu \in \mathcal{M}_{\epsilon_1, n}^* \mid \mu \xleftarrow{R} M_n^*] > 1 - \epsilon_2$. Therefore, $\forall \epsilon_1, \epsilon_2 > 0$, $\exists n_0 \in \mathbb{N}$, $\#\mathcal{M}_{\epsilon_1, n}^* > 2^{H(M_n^*) - \epsilon_2}$.

Due to the definition of conditional semantic information amount $\text{SA}(X(\mu) : \mathcal{Z}, \delta(n))$, there exist at most $2^{\text{SA}(X(\mu) : \mathcal{Z}, \delta(n))}$ distinct values of $\mu^- \in \mathbb{M}_n^{\overline{\mathcal{C}}^{\mathcal{X} : \mathcal{Z}}}$ such that the output of $\text{Ch}(n, X(\mu))$ is indistinguishable from that of $\text{Ch}(n, X(\mu^-))$. Let $\mathcal{I}_\mu \subseteq \mathbb{M}_n^{\overline{\mathcal{C}}^{\mathcal{X} : \mathcal{Z}}}$ be the set of such at most $2^{\text{SA}(X(\mu) : \mathcal{Z}, \delta(n))}$ values of μ^- .

If $\mathcal{I}_\mu \cap \mathcal{I}_{\mu'} \neq \lambda$ for $\mu \neq \mu'$ (the intersection of the two sets is not empty), there exists $\mu^- \in \mathcal{I}_\mu \cap \mathcal{I}_{\mu'}$ such that the output of $\text{Ch}(n, X(\mu^-))$ is indistinguishable from both $\text{Ch}(n, X(\mu))$ and $\text{Ch}(n, X(\mu'))$, i.e., the output of $\text{Ch}(n, X(\mu))$ is indistinguishable from that of $\text{Ch}(n, X(\mu'))$. That is, $\mathcal{I}_\mu = \mathcal{I}_{\mu'}$. Therefore, for $\mu \neq \mu'$, either $\mathcal{I}_\mu \cap \mathcal{I}_{\mu'} = \lambda$ or $\mathcal{I}_\mu = \mathcal{I}_{\mu'}$. Hence, we have t disjoint sets $\mathcal{I}_{\mu_1}, \mathcal{I}_{\mu_2}, \dots, \mathcal{I}_{\mu_t}$ for some $t \in \mathbb{N}$. Therefore, from the above property, $\mathcal{M}_{\epsilon_1, n}^*$ is divided into disjoint equivalence classes, $\mathcal{I}_i^* := \mathcal{I}_{\mu_i} \cap \mathcal{M}_{\epsilon_1, n}^*$, $i = 1, \dots, t$.

Since $|\text{SA}(X(\mu_i) : \mathcal{Z}, \delta(n)) - T^*| \leq \epsilon_1$ for $\mu_i \in \mathcal{M}_{\epsilon_1, n}^*$, we obtain $\#\mathcal{I}_{\mu_i} \leq 2^{\text{SA}(X(\mu_i) : \mathcal{Z}, \delta(n))} \leq 2^{T^* + \epsilon_1}$. Hence, $t \geq \#\mathcal{M}_{\epsilon_1, n}^* / \max\{\#\mathcal{I}_{\mu_i}\} \geq 2^{H(M_n^*) - \epsilon_1} / 2^{T^* + \epsilon_1} = 2^{H(M_n^*) - T^* - 2\epsilon_1}$. Let $t^* := 2^{H(M_n^*) - T^* - 2\epsilon_1}$.

We now set a coding with n , $\text{code}_n : \{0, 1\}^{k(n)} \rightarrow \mathcal{M}_n^{\overline{\mathcal{C}}}$, such that $\text{code}_n : i \mapsto \mu_i \in \mathcal{I}_i^*$ for $i = 1, \dots, t^*$. That is, $k(n) := \log t^* = H(M_n^*) - T^* - 2\epsilon_1$, i.e., $\text{SR}(\text{code}_n) := k(n)/n = (H(M_n^*) - T^* - 2\epsilon_1)/n = \text{SC}(\text{SChC}_n, \delta(n)) - 2\epsilon_1/n$, i.e., $\text{SR}(\text{code}_n) = \text{SC}(\text{SChC}_n, \delta(n)) - 2\epsilon_1/n$.

Since $\mathcal{I}_i^* (i = 1, \dots, t^*)$ are disjoint and $\text{Ch}(n, X(\mu_i))(i = 1, \dots, t^*)$ are distinct, the coding by code_n can be uniquely decoded.

Thus, for any ϵ there exists n_0 such that for all $n > n_0$ the coding system satisfies $\text{SC}(\text{SChC}_n, \delta(n)) - \epsilon < \text{SR}(\text{code}_n) < \text{SC}(\text{SChC}_n, \delta(n))$. This completes the proof of this theorem. \square

Remark 14 The error correction techniques based on the traditional (Shannon) information theory have been widely used in many applications, but they are incompetent for correcting various types of semantic errors which are described in [2].

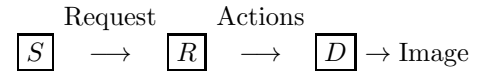
In this paper, the notion of a semantic channel coding system or semantic error correction is introduced. Theorem 7 shows a great potential of semantic error correction techniques. For example, we can correct semantic errors in a natural language sentence using the semantics and context. Such a capability of human beings is theoretically captured and formalized in Theorem 7. Unfortunately, the proof of the theorem ignores the efficiency and the error-correction technique used in the proof is impractical, i.e., it only gives a theoretical feasibility. However, such a feasibility result indicated by Theorem 7 should push toward developing practical techniques. This is similar to that where Shannon's channel coding theorem is only a feasibility result and many practical error correcting codes have been developed which are quite different

from Shannon's random coding technique. Towards practical semantic error correcting code techniques, artificial intelligence technologies might offer some potent means to yield a breakthrough in this field.

4. Effectiveness Problem

We now consider the problem of effectiveness (Level C problem given by Weaver [41] as introduced in Section IA) in our semantic information theory.

First consider the following experiment. Person S requests robot R , e.g., by voice using English sentences, to perform a series of actions, and device D detects the image of the actions of robot R and outputs the (digital form of) video image.



In the experiment, then, an evaluator, e.g., a human, E compares the request and the image, and evaluates how correctly the robot understood the request and performed the requested actions.

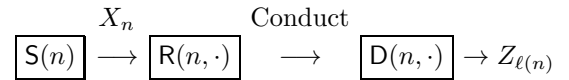
The problems here are that the request and image are different types of information, e.g., the request is the (digital form) voice speaking English sentences and the image is the (digital form) video images of the robot's actions. Although they are different forms of information, human E can evaluate the capability of robot R since E knows some semantics common between English requests and the robot's actions.

Our semantic information theory can treat such semantics that are common between different forms.

Definition 15 (Effectiveness)

Message-to-conduct system $\text{M2C}(S, R, D; \mathcal{X}, \mathcal{Z})$ consists of sender S and receiver R where S sends semantic information source \mathcal{X} to R and R 's conduct is observed by some device, D , which outputs a sequential family of distributions, \mathcal{Z} .

For parameter $n \in \mathbb{N}$, message $X_n \in \mathcal{X}$, which is the distribution over $\{0, 1\}^n$, is given to R and R 's conduct is observed as distribution $Z_{\ell(n)}$, which is the distribution over $\{0, 1\}^{\ell(n)}$.



If \mathcal{X} is a conditional semantic information source under \mathcal{Z} , we can define the concept of "effectiveness" as follows:

The effectiveness, $\text{Eff}(X_n : \mathcal{Z}, \delta(n))$, of message-to-conduct system $\text{M2C}(S, R, D; \mathcal{X}, \mathcal{Z})$ is

$$\text{Eff}(X_n : \mathcal{Z}, \delta(n)) := \frac{\text{SI}(X_n : \mathcal{Z}, \delta(n))}{\text{SA}(X_n, \delta(n))}.$$

The effectiveness represents the ratio of how much portion of semantics of original message \mathcal{X} is understood and conducted correctly by R .

In the above-mentioned experiment, if robot R is very capable, R should correctly recognize various requests from S and act accordingly as requested. For example, R recognizes a million ($\approx 2^{20}$) requests and behaves correctly as requested. In this case, we can consider that the effectiveness complexity is approximately 20 bits (similar to 20 bits of a semantic channel rate). On the other hand, if R is not so capable, for example, R recognizes only 8 requests and behaves correctly, then its effectiveness complexity is only 3 bits.

In this example, the requests by S are formalized as elements of semantic information space $\mathbb{M}^{\overline{\mathcal{C}}}$, and (R, D) is considered as a functionality similar to a semantic channel in Section IIIB3, where various semantic errors occur. Hence, the capability of robot R (effectiveness complexity) with a distribution of instructions, and the error correcting functionality in the system are characterized by the *effectiveness capacity* of (R, D) and *effectiveness rate* of a semantic coding, which are defined in a manner similar to that of semantic channel capacity and semantic communication rate (Definition 13), respectively. Roughly, the effectiveness capacity indicates the rate of the maximum capability of robot R (effectiveness complexity) and the effectiveness rate is the communication rate of semantic (error correcting) coding.

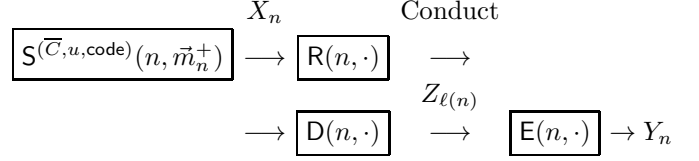
First we introduce the notion of an *effectiveness coding system* in a manner similar to that of the semantic channel coding system.

Definition 16 (*Effectiveness Coding System*)

Effectiveness coding system $\text{ECS} := \text{ECS}(\mathcal{S}^{\overline{\mathcal{C}}, u, \text{code}}, R, D, E; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ consists of sender S , receiver R , device D and evaluator E . Here, S has a coding machine, $(\overline{\mathcal{C}}, u, \text{code})$, and sends a sequential family of distributions, \mathcal{X} to R . Receiver R then performs actions given \mathcal{X} , and R 's conduct is observed by device D , which outputs a sequential family of distributions, \mathcal{Z} , and sends it to E . Evaluator E is a conditional oc-circuit under \mathcal{Z} and outputs \mathcal{Y} .

Here, \mathcal{X} is generated by an oc-circuit, $\mathcal{C} := (\overline{\mathcal{C}}, u, \infty, \vec{m}_\infty)$, where $\overline{\mathcal{C}} := (C, N_u, N_s, N_m, N_r, L_y, s_1)$ (logic), $u \in \{0, 1\}^{N_u}$ (universe space), $\vec{m}_\infty \in \mathbb{M}^{\overline{\mathcal{C}}} := \{(m_i)_{i=1,2,\dots} \mid m_i \in \{0, 1\}^{N_m}\}$ (semantic information space), and $\mathbb{M}_n^{\overline{\mathcal{C}}} = \{0, 1\}^{N_m K_n}$ (the $N_m K_n$ -bit prefix of $\mathbb{M}^{\overline{\mathcal{C}}}$ for n -bit output) ($K_n := \lceil n/L_y \rceil$).

For parameter $n \in \mathbb{N}$, given $\vec{m}_n^+ \in \{0, 1\}^{k(n)}$, S computes $\vec{m}_n := \text{code}_n(\vec{m}_n^+) \in \mathbb{M}_n^{\overline{\mathcal{C}}}$ and $X_n \stackrel{R}{\leftarrow} (\overline{\mathcal{C}}, u, n, \vec{m}_n)$, where $X_n \in \mathcal{X}$ is a distribution over $\{0, 1\}^n$. X_n is input to receiver R , and R 's conduct is observed by device D , which outputs $Z_{\ell(n)} \in \mathcal{Z}$, where $Z_{\ell(n)}$ is a distribution over $\{0, 1\}^{\ell(n)}$. Evaluator E (conditional oc-circuit under \mathcal{Z}) receives $Z_{\ell(n)}$ and outputs $Y_n \in \mathcal{Y}$, where Y_n is a distribution over $\{0, 1\}^n$.



We say that $\text{ECS} := \text{ECS}(\mathcal{S}^{\overline{\mathcal{C}}, u, \text{code}}, R, D, E; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ correctly codes at precision level $\delta(\cdot)$ if there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $X_n \stackrel{\delta(n)}{\approx} Y_n$.

Definition 17 (*Effectiveness Capacity and Effectiveness Rate*)

Let $\text{ES} := (\mathcal{S}^{\overline{\mathcal{C}}, u}, R, D; \mathcal{X}, \mathcal{Z})$ be a “*effectiveness system*” of effectiveness coding system $\text{ECS}(\mathcal{S}^{\overline{\mathcal{C}}, u, \text{code}}, R, D, E; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$. Given $\vec{m}_\infty \in \mathbb{M}^{\overline{\mathcal{C}}}$, S in ES computes $\mathcal{X} := \{X(\vec{m}_n) := X_n \stackrel{R}{\leftarrow} (\overline{\mathcal{C}}, u, n, \vec{m}_n)\}_{n \in \mathbb{N}}$, i.e., $\mathcal{X} \stackrel{R}{\leftarrow} (\overline{\mathcal{C}}, u, \infty, \vec{m}_\infty)$, where \vec{m}_n is the $N_m \cdot \lceil n/L_y \rceil$ -bit prefix of \vec{m}_∞ for n -bit output. Then, $\mathcal{X} := \{X(\vec{m}_n)\}_{n \in \mathbb{N}}$ is input to R in ES , and D outputs $\mathcal{Z} := \{Z_{\ell(n)}\}_{n \in \mathbb{N}}$.

If for any $\vec{m}_\infty \in \mathbb{M}^{\overline{\mathcal{C}}}$, $\mathcal{X} \stackrel{R}{\leftarrow} (\overline{\mathcal{C}}, u, \infty, \vec{m}_\infty)$ is a conditional semantic information source under \mathcal{Z} in effectiveness system ES , we call ES “*normal*”.

Let $\mathcal{M} := \{M_n\}_{n \in \mathbb{N}}$ be a sequential family of distributions over $\mathbb{M}^{\overline{\mathcal{C}}}$, where M_n is a distribution over $\mathbb{M}_n^{\overline{\mathcal{C}}}$, $n \in \mathbb{N}$, i.e., $M_n := \{(\vec{m}_n, p_{\vec{m}_n}) \mid \vec{m}_n \in \mathbb{M}_n^{\overline{\mathcal{C}}}\}$, (see Definition 5 for the sequential family of distributions).

When effectiveness system ES is normal, “*effectiveness capacity*” EC of ES for $n \in \mathbb{N}$ (say ES_n) is

$$\text{EC}(\text{ES}_n, \delta(n)) := \frac{1}{n} \cdot \max_{M_n \in \mathcal{M}_n^{\text{seq}}} \{H(M_n) - \mathbb{E}_{M_n}(\text{SA}(X(\vec{m}_n) : \mathcal{Z}, \delta(n)))\}, \quad (6)$$

where $\mathcal{M}_n^{\text{seq}}$ is the class of the $N_m \cdot \lceil n/L_y \rceil$ -bit prefix of sequential families of distributions, i.e., $M_n \in \mathcal{M}_n^{\text{seq}}$ is the $N_m \cdot \lceil n/L_y \rceil$ -bit prefix of a sequential family of distributions, $H(\cdot)$ is the Shannon entropy and $\mathbb{E}_{M_n}(\cdot)$ is the expectation value over the distribution of $\vec{m}_n \stackrel{R}{\leftarrow} M_n \in \mathcal{M}_n^{\text{seq}}$.

“*Effectiveness rate*” ER of effectiveness coding $(\overline{\mathcal{C}}, u, \text{code})$ in effectiveness coding system $\text{ECS}(\mathcal{S}^{\overline{\mathcal{C}}, u, \text{code}}, R, D, E; \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ for $n \in \mathbb{N}$ is

$$\text{ER}(\text{code}_n) := k(n)/n. \quad (7)$$

We say a effectiveness coding system, ECS , is “*normal*,” if the effectiveness system, ES , of ECS is normal.

We can also define the “*uniformity*” of normal effectiveness coding system ECS in the same manner as that for the normal semantic channel coding system described in Definition 14.

Theorem 8 (*Effectiveness Coding Theorem*)

There exists a uniform effectiveness coding system, ECS, that correctly codes at precision level $\delta(\cdot)$ and for any ϵ ($0 < \epsilon$), there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$\text{EC}(\text{ES}_n, \delta(n)) - \epsilon < \text{ER}(\text{code}_n) < \text{EC}(\text{ES}_n, \delta(n)).$$

There exists no normal effectiveness coding system, ECS, that correctly codes at precision level $\delta(\cdot)$ (with a negligible error probability) and for any ϵ ($0 < \epsilon$), there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$\text{EC}(\text{ES}_n, \delta(n)) < \text{ER}(\text{code}_n).$$

IV. CONCLUSION

Approximately seven decades have passed since Warren Weaver published his two insightful and prescient articles [40, 41] that clearly indicated two research directions in science, organized complexity and semantic information theory. Although the articles stimulated and encouraged these research areas, it is hard to say that these areas have been well established in science, and Weaver would be disappointed to know it.

Moreover, he might be disappointed to learn that no study has been done on the relation and integration of these areas, since he could have realized the relationship between the areas considering that these articles were written at almost the same time.

The aim of this paper is to pursue the research directions that Weaver indicated. This paper first quantitatively defined the organized complexity. The proposed definition for the first time simultaneously captures the three major features of organized complexity and satisfies all of the criteria for organized complexity measures introduced in this paper. We then applied the organized complexity measure to develop our semantic information theory, where we presented the first formal definition of a semantic information amount that is based only on concretely defined notions, and unveil several fundamental properties in the semantic information theory. Through this organized complexity measure, we offered a unified paradigm of organized complexity and semantic information theory.

Organized complexity is an interdisciplinary concept straddling physics, cosmology, biology, ecology, sociology, and informatics. Thus, the proposed organized complexity measure could be a core notion in such interdisciplinary areas, and for example, offer some basis for tackling the problems posed in [24].

-
- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle, Proceedings of the 2nd International Symposium on Information Theory, Petrov, B. N., and Caski, F. (eds.), Akademiai Kiado, Budapest: 267-281 (1973).
 - [2] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler. Towards a theory of semantic communication, IEEE International Workshop on Network Science (2011).
 - [3] F. Bacchus. On probability distributions over possible worlds. In UAI, pp. 217–226, (1988).
 - [4] C. H. Bennett. Logical depth and physical complexity. In R. Herken, editor, The Universal Turing Machine, A Half-Century Survey, pages 227–257. Oxford University Press, Oxford, (1988).
 - [5] R. Carnap, and Y. Bar-Hillel. An outline of a theory of semantic information. RLE Technical Reports 247, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge MA, Oct (1952).
 - [6] G. J. Chaitin. On the length of programs for computing finite binary sequences: statistical considerations. J. Assoc. Comput. Mach., 16, pp.145–159 (1969).
 - [7] T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley & Sons (1991).
 - [8] J. P. Crutchfield and K. Young. Inferring statistical complexity. Physical Review Letters, 63:105–108, (1989).
 - [9] J. P. Crutchfield. The calculi of emergence: computation, dynamics and induction. Physica D: Nonlinear Phenomena, 75(1-3):11–54, (1994).
 - [10] J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: objective complexity via minimal representations. Physical Review E, 59(1):275–283 (1999)
 - [11] S. D’Alfonso, On quantifying semantic information. Information 2, 1, 61.101 (2011).
 - [12] H. B. Enderton. A Mathematical Introduction to Logic S. 2nd ed., Academic Press (2011).
 - [13] L. Floridi. Outline of a theory of strongly semantic information. Minds Mach. 14, 2, 197.221 (2004).
 - [14] L. Floridi. Philosophical conceptions of information. In Sommaruga [38], pp. 13–53 (2009).
 - [15] M. Gell-Mann. What is complexity. Complexity, 1:1 (1995)
 - [16] M. Gell-Mann and S. Lloyd. Information measures, effective complexity, and total information. Complexity, 2(1):44–52 (1996)
 - [17] M. Gell-Mann and S. Lloyd. Effective complexity. In M. Gell-Mann and C. Tsallis, editors, Nonextensive Entropy Interdisciplinary Applications. The Santa Fe Institute, OUP USA (2004).
 - [18] P. Grassberger. Problems in quantifying self-generated complexity. Helvetica Physica Acta, 62: 489–511 (1989)
 - [19] B. Juba, and M. Sudan. Universal semantic communication i. In STOC, pp. 123–132 (2008).
 - [20] B. Juba, and M. Sudan. Universal semantic communication ii: A theory of goal-oriented communication. Electronic Colloquium on Computational Complexity (ECCC) 15, 095 (2008).
 - [21] J. Kohlas, and C. Schnewly. Information algebra. In Sommaruga [38], pp. 95–127 (2009).
 - [22] A. N. Kolmogorov. Three approaches to the quantitative definition of information. Problems Inform. Trans-

- mission, 1(1), pp.1–7 (1965).
- [23] J. Langel. Logic and Information, A Unifying Approach to Semantic Information Theory. Ph.d. dissertation, Universitat Freiburg in der Schweiz (2009).
 - [24] C. H. Lineweaver, P. C. W. Davies, M. Ruse, (eds.). Complexity and the Arrow of Time, Cambridge University Press (2013).
 - [25] J. Ladyman, J. Lambert, and K. Wiesner. What is a complex system?, *European Journal for Philosophy of Science* 3, no. 1 pp. 33–67, (2013).
 - [26] A. Lempel, and J. Ziv. Compression of twodimensional data. *IEEE Transactions in Information Theory*, IT-32:2–8 (1986).
 - [27] M. Li, and P. M. B. Vitanyi. An Introduction to Kolmogorov Complexity and Its Applications. Graduate Texts in Computer Science. Springer Verlag, 2nd edition (1997).
 - [28] S. Lloyd, and H. Pagels. Complexity as thermodynamic depth. *Annals of Physics*, 188:186–213, 36 (1988).
 - [29] J. M. D. Nafria, and F. S. Alemany, Eds. What is really information? An interdisciplinary approach (2009), vol. 7, tripleC.
 - [30] M. A. Nielsen, and I. L. Chuang. Quantum computation and quantum information, Cambridge University Press (2000).
 - [31] N. J. Nilsson. Probabilistic logic. *Artif. Intell.* 28, 1, 71.87 (1986).
 - [32] J. Rissanen. Stochastic Complexity in Statistical Inquiry. World Scientific, Singapore (1989).
 - [33] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions in Information Theory*, IT-30:629–636 (1984).
 - [34] M. Sipser. Introduction to the Theory of Computation, Third Edition, Cengage Learning (2013).
 - [35] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423; 623–656 (1948).
 - [36] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3):817–879 (2001).
 - [37] R. J. Solomonoff. A formal theory of inductive inference, part 1 and part 2. *Inform. Contr.*, 7, pp.1–22, pp.224254 (1964).
 - [38] G. Sommagura, Ed. Formal Theories of Information: From Shannon to Semantic Information Theory and General Concepts of Information [Muenchenwiler Seminar (Switzerland), May 2009], vol. 5363 of Lecture Notes in Computer Science, Springer (2009).
 - [39] H. Vollmer. Introduction to Circuit Complexity: a Uniform Approach. Springer Verlag (1999).
 - [40] W. Weaver. Science and Complexity. *American Scientist* 36 (4): 536–544 (1948).
 - [41] W. Weaver. The Mathematical Theory of Communication, ch. Recent Contributions to the Mathematical Theory of Communication, Univ. of Illinois Press (1949).
 - [42] F. M. Willems, and T. Kalker. Semantic compaction, transmission, and compression codes. In *Proceedings of International Symposium on Information Theory (ISIT)*, pp. 214–218 (2005).
 - [43] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions in Information Theory*, IT-23:337–343 (1977).